UNIVERSITE DE TOULON ET DU VAR

N° attribué par la bibliothèque
|_|_|_|_|_|_|_|_|_|_|_|_|

THESE POUR L'OBTENTION DU DOCTORAT EN SCIENCES DE L'INFORMATION ET DE LA COMMUNICATION A L'UNIVERSITE DE TOULON ET DU VAR
conforme au nouveau régime défini par l'arrêté du 30 mars 1992

METHODOLOGIE ET STRUCTURATION D'UN OUTIL DE DECOUVERTE DE CONNAISSANCES BASE SUR LA LITTERATURE BIOMEDICALE : UNE APPLICATION BASEE SUR L'EXPLOITATION DU MESH

présentée et soutenue publiquement le 28 février 2006

par Jean-Dominique PIERRET

Galderma R&D Sophia-Antipolis

Sous la Direction de
M. Luc QUONIAM
Professeur à l'Université de Toulon et du Var

Membres du jury :

M. Yves-François LE COADIC (Rapporteur)
Professeur au Conservatoire National des Arts & Métiers

M. Thierry LAFOUGE (Rapporteur)
Professeur à l'Université Claude Bernard

M. Fabrizio DOLFI
Docteur en Médecine, Galderma R&D Sophia-Antipolis

M. Eric BOUTIN (Tuteur)
Maître de conférences à l'Université de Toulon et du Var

Je tiens à remercier la société Galderma R&D Sophia-Antipolis pour m'avoir permis de réaliser ce travail, mais plus encore, pour avoir réuni les conditions nécessaires à la genèse du DPM.

Merci à Luc Quoniam de m'avoir transmis le virus de la bibliométrie, voici maintenant plus de 13 ans, et cette idée que derrière une fréquence, aussi faible soit-elle, peut se cacher une information d'une grande valeur. Je suis très honoré que Luc soit mon Directeur de thèse.

Merci à Eric Boutin pour son support constant et enthousiaste tout au long de ce travail. L'idée de réaliser cette thèse sur le DPM revient à Eric, je le remercie pour cette initiative.

Sans Fabrizio Dolfi, critique dynamique, constructif et éclairé, le DPM n'aurait certainement pas existé. Merci à toi mon ami.

Merci aux professeurs Yves-François Le Coadic et Thierry Lafouge pour avoir accepté d'être rapporteurs de cette thèse.

Merci enfin à celles et ceux qui ont contribué à la réalisation de ce projet : Nadège Tremel, Annick Pierret, Marie-Jo Lejard, Irina Safonova, Christian Gerini, Christian Loesche, Philippe Walter et Marc Weeber.

A Nadège

The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' (I found it!) but 'That's funny ...'

Isaac Asimov

SOMMAIRE

Index des figures

Index des tableaux

ABREVIATIONS ET CONVENTIONS D'ECRITURE

DPM : Diseases Physiopathology Molecules

KDD : Knowledge Discovery in Database

MeSH : Medical Subject Headings (thesaurus de la NLM)

NCBI : National Center for Biotechnology Information

NIH : National Institute of Health

NLM : National Library of Medicine

UMLS : Unified Medical Language System

EPA : Eicosapentaenoic Acid

Référence bibliographique : citation

Afin de simplifier la lecture de la bibliographie de ce manuscrit, nous l'avons séparé en deux parties. La bibliographie générale, dont les références figurent en police normale [Swanson, 1986b], rapportée dans la section bibliographie. La bibliographie sur laquelle Don Swanson s'est appuyé pour ses travaux, dont les références figurent en italique *[Dyerberg, 1982]* et sont présentées dans l'annexe 1. Nous rappelons cette convention en début de première partie.

Nous emploierons de nombreux descripteurs du MeSH. Convenons dès à présent de les représenter par la police de caractère `courier new`. Les requêtes PubMed seront également ainsi représentées.

Le vrai chercheur doit savoir faire attention aux signes qui révéleront l'existence d'un phénomène auquel il ne s'attend pas.

Louis Leprince-Ringuet

INTRODUCTION

Les professionnels de l'information biomédicale exercent dans un environnement où les bases de données bibliographiques sont nombreuses et généralement très bien structurées, employant codes (molécules, gènes, protéines) et thesaurus. Le volume d'information qu'elles mettent à leur disposition est conséquent. Pour situer le contexte dans lequel a été réalisé ce travail, nous présentons ici brièvement les bases de données que nous utilisons dans notre pratique quotidienne :

- Medline, produite par la *National Library of Medicine*, comptant à ce jour plus de 13 millions de citations. Chaque citation est indexée avec les descripteurs du MeSH, thesaurus développé par la NLM. Medline est accessible gratuitement par l'interface PubMed mais également par des serveurs commerciaux (DataStar, STN, Dialog, …). Medline/PubMed et le MeSH sont les sources que nous utilisons pour le présent travail.
- Embase[1], produite par Elsevier Science BV (Amsterdam), comptant près de 10 millions de citations indexées par le thesaurus EMTREE.
- SciSearch, dont la particularité est de permettre la recherche par référence citée. Elle est produite par Thomson ISI.
- Biosis, produite par Thomson ISI, comptant plus de 13 millions de citations.

Embase, SciSearch et Biosis sont accessibles par des serveurs commerciaux (DataStar, STN, Dialog …). D'autres sources, comme Pascal, les Derwent Drug Files ou ToxLine sont également employées.

Augmentation du volume d'informations…

La requête Medline `"Acne Vulgaris/drug therapy"[MAJR]`, a donné 1.937 références bibliographiques le 23 janvier 2005. Il est impossible de traiter manuellement ce volume d'information ; il convient de restreindre la recherche de différentes manières (par la date ou en ajoutant d'autres critères).

L'interface PubMed connaît une forte croissance de son volume. Ainsi, toujours le 23 janvier 2005, PubMed contenait 15.341.735 citations. Un peu plus de 623.000 nouvelles citations ont été ajoutées annuellement depuis 2001.

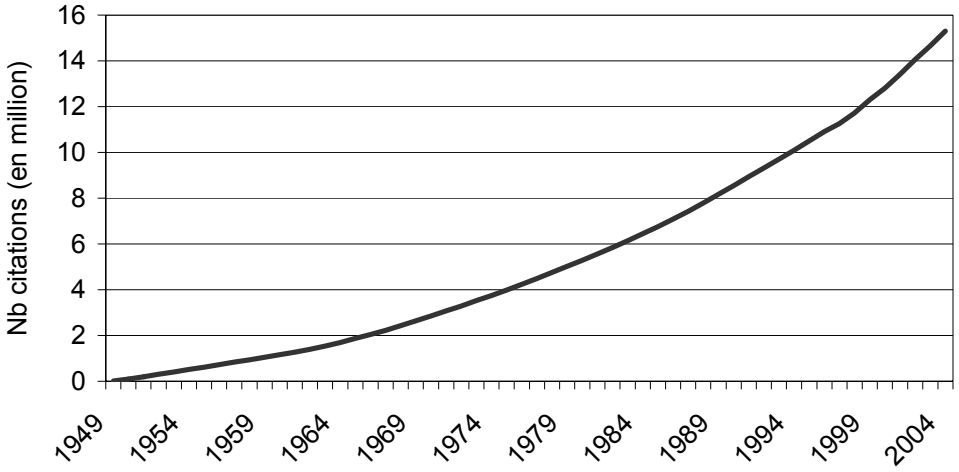


Figure 1 : Evolution du volume de citations de PubMed

La figure 1 illustre la croissance du nombre total de citations biomédicales de PubMed, reprenant les sources qu'elle compile. Nous parlons ici de PubMed au

[1] Elsevier B.V. (Page consultée le 22 septembre 2005). *Embase suite of products*, [En ligne]. Adresse URL : http://www.info.embase.com/embase_suite

sens large, incluant OLDMedline[2], Medline, les citations en cours de traitement et les citations des articles de PubMedCentral[3].

De la même manière, le nombre de bases de données publiques dans le domaine des sciences de la vie était estimé à un peu plus de 50 en 1992 et à plus de 450 en 2002 [Wren, 2004a].

La quantité d'information générée à travers le monde s'accroît de manière formidable : ainsi, considérant toute l'information disponible stockée sur papier, film, supports magnétiques ou optiques, 5 exabytes[4] d'informations ont été produits en 2002 selon le rapport "*How much information ? 2003*". Cette quantité a augmenté de 30% par an entre 1999 et 2002 [Lyman, 2003].

Paradoxalement, malgré cette abondance, il est difficile de trouver la bonne information [Edmunds 2000]. Grivell écrivait en 2002 "*amongst the many prophesies of doom, relatively little attention has been paid to the consequences of the growing amount of scientific literature*" [Grivell, 2002]. Dans les sciences de la vie, cette surcharge n'est pas sans poser de problèmes [Fogarty, 2002]. Ne citons que le cas du décès d'une volontaire de 24 ans participant à une étude sur l'asthme en juin 2001 à l'université Johns Hopkins. Cette affaire a eu un large écho dans le milieu de l'information médicale. Lors de cette étude, les volontaires inhalaient de l'hexaméthonium, substance responsable du décès d'Ellen Roche. Le médecin qui supervisait l'étude a conduit ses recherches bibliographiques sur la toxicité de l'hexaméthonium en se basant sur PubMed. A l'époque, PubMed couvrait très partiellement la période antérieure à 1966. Se contentant de cette unique source, il n'a pu identifier les articles ou données factuelles, pourtant présentes dans d'autres bases de données (Toxline[5] ou

[2] OLDMedline : Medline a été crée en 1966, OLDMedline contient les citations antérieures à cette date.

[3] Archive numérique de revues en sciences de la vie développée et gèrée par le *National Center for Biotechnology Information* (NCBI, Bethesda, Mariland),), accessible par l'interface PubMed. NCBI, National Library of Medicine. (Page consultée le 22 septembre 2005). *PubMed*, [En ligne]. Adresse URL : http://www.ncbi.nlm.nih.gov

[4] $5\ 10^{18}$ bytes

[5] Toxline : base de données bibliographique accessible librement par l'interface Toxnet de la NLM. Toxline contient environ 3 millions de citations sur les effets biochimiques, pharmacologiques, physiologiques et toxicologiques de médicaments ou substances chimiques.

Poisindex[6]) et qui mettent en évidence les risques liés à l'usage de ce produit. La plupart des articles relatifs à ce drame expliquent qu'il aurait pu être évité si la recherche bibliographique avait été conduite par un documentaliste spécialisé [Josefson, 2001], [Marshall, 2001], [McCarthy, 2001], [McLellan, 2001], [Perkins, 2001] et [Smaglik, 2001].

Pour illustrer la préoccupation grandissante des scientifiques face à la croissance de la quantité de données disponibles, nous avons interrogé la base bibliographique SciSearch avec la requête "`(needle OR needles) AND (haystack OR haystacks)`" le 23 janvier 2005. Nous prenons le risque de l'approximation suivante : quand une citation contient à la fois *needle* et *haystack*, il est hautement probable qu'elle aborde la question de la recherche d'information dans un contexte de surcharge de données (voir par exemple [Grivell, 2002]). La figure 2 illustre la répartition du nombre de publications au cours du temps.

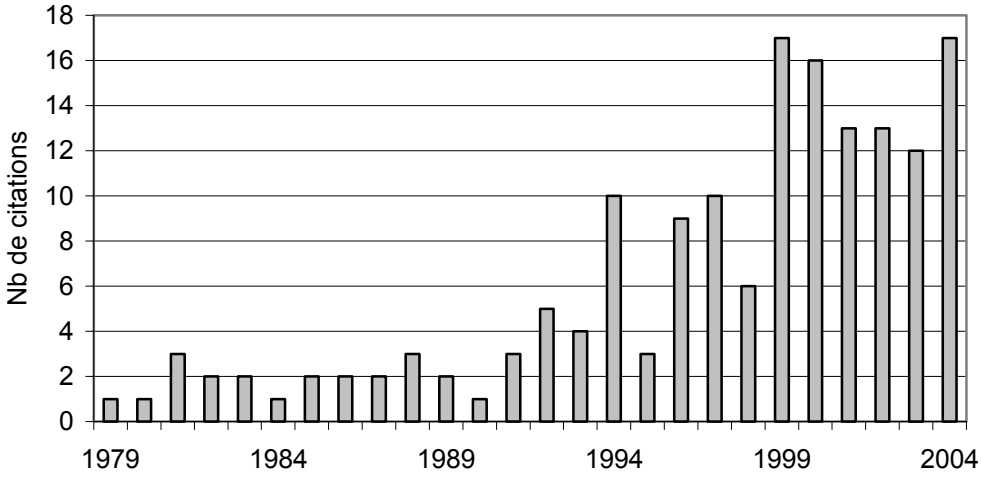


Figure 2 : Répartition annuelle du nombre de citations SciSearch contenant les mots *needle* et *haystack*

National Library of Medicine. (Page consultée le 22 septembre 2005). Toxnet, Toxicologic Data Network. [En ligne]. Adresse URL : http://toxnet.nlm.nih.gov

[6] Poisindex : base de données factuelle contenant des informations sur la conduite à tenir en cas d'exposition toxique. Poisindex est un produit de Micromedex. Thomson. (Page consultée le 22 septembre 2005). *Thomson Micromedex*, [En ligne]. Adresse URL : http://www.micromedex.com

Sans entrer dans une analyse complexe, nous pouvons sans risque dire que la problématique du trop-plein d'information fait couler de plus en plus d'encre.

…et fragmentation du savoir…

Le temps est loin où l'*Encyclopédie ou dictionnaire raisonné des sciences, des arts et des métiers* de Diderot et d'Alembert permettait à l'érudit d'appréhender l'ensemble de la connaissance - en 28 volumes. Aujourd'hui, nos connaissances grandissent sans cesse et la littérature scientifique augmente en conséquence. C'est, en première approximation, la résultante de la croissance du nombre de chercheurs et savants. Afin de pouvoir traiter, *digérer*, ce volume en pleine expansion, la communauté scientifique semble s'organiser en se partageant le travail, en identifiant des problématiques de plus en plus précises, en créant des disciplines, des spécialités toujours plus pointues [Swanson, 1993]. Par un mécanisme obscure, les scientifiques, qui ont depuis longtemps abandonné l'idée de pouvoir lire tout les écrits qui paraissent, se sont organisés pour travailler en spécialités, permettant à chacun de se concentrer sur une petite partie de la littérature. Les spécialités qui croissent trop vite se divisent à leur tour en "sous spécialités". Ainsi le volume d'information afférent à ces nouvelles spécialités est à peu près constant et gérable par les spécialistes. Au fur et à mesure que croit la littérature, le nombre de spécialités augmente, laissant pour chacune d'entre elles un volume approximativement constant d'articles. Ainsi, le rapport entre le nombre de savants et la littérature reste constant, permettant à chacun de suivre – ou laissant à chacun l'illusion de suivre – la littérature de son domaine. La conséquence de cette augmentation de spécialités est la fragmentation des connaissances. En divisant ainsi la production scientifique, les relations logiques entre différentes spécialités tendront à être négligées, ignorées ou occultées. Or, deux spécialités peuvent sur certains points être complémentaires et porteuses, ensemble, de nouveaux savoirs. Car même si les scientifiques se sont organisés de manière à suivre ce qui est produit dans leurs champs d'expertise – de plus en plus restreints – il est aujourd'hui impossible à un seul homme d'embrasser l'ensemble des

connaissances d'un domaine tel que la biologie moléculaire ou la physique quantique. L'éloignement de disciplines qui peuvent être complémentaires à certains égards et l'augmentation du volume du savoir, contribuent à accroître le nombre de connexions porteuses de nouvelles connaissances.

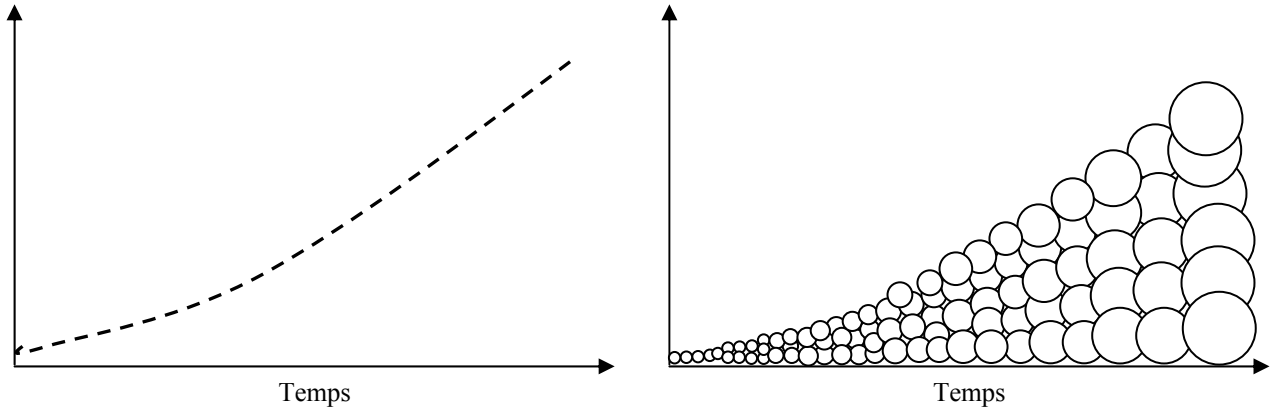


Figure 3 : Evolution du volume d'information disponible au cours du temps

Les deux graphes de la figure 3 (réalisés d'après Swanson) illustrent comment augmente le volume d'informations. Celui de gauche reflète une augmentation d'un point de vue macroscopique : au cours du temps, plus d'informations et de connaissances sont produites. Celui de droite illustre le même phénomène mais rend mieux compte de la fragmentation des connaissances au fur et à mesure qu'elles s'accroissent, qui se répartissent en sphères de recherche, en thématiques bien isolées avec de moins en moins de liens directes entre elles.

La fragmentation du savoir cache ces relations pertinentes. Le travail de Don Swanson, à la base de notre méthode de découverte de connaissances, porte sur l'identification de ces relations à travers l'exploitation des bases de données bibliographiques. L'expression qu'il emploie est *undiscovered public knowledge* [Swanson, 1986a]. Ces connexions sont révélées par une méthode originale d'exploitation des bases de données bibliographiques suivant le modèle transitif de Swanson. Le volume des bases de données étant en expansion constante, les moyens à appliquer doivent être capable de traiter massivement de grands volumes d'information. Le travail présenté dans ce mémoire s'intéresse à la découverte de ces connexions cachées, porteuses de nouveaux savoirs.

…vers un nouveau mode d'exploitation des bases de données bibliographiques

Le concept de "découverte de connaissances dans les bases de données bibliographiques" comporte donc un paradoxe. Traditionnellement, en Sciences de l'Information, on considère que l'information disponible dans les bases de données bibliographiques est une information datée, validée par un processus long qui la rend peu innovante. D'ailleurs, par construction, les bases de données sont classiquement interrogées de manière booléenne : le résultat d'une requête est un ensemble d'informations connues qui n'apporte en lui-même aucune nouveauté, on obtient ce que l'on a recherché.

Cette thèse s'inscrit donc dans la démarche amorcée par Don Swanson. Ses travaux montrent le potentiel insoupçonné des bases bibliographiques dans la révélation et la découverte de connaissances. Cet intérêt ne tient pas tant à la nature de l'information disponible qu'à la méthodologie utilisée pour révéler ces nouvelles connaissances. Cette méthodologie générale s'applique de façon privilégiée dans un environnement d'information validée et structurée ce qui est le cas de l'information bibliographique.

L'expression "découverte de connaissances dans les bases de données" est la traduction de *Knowledge Discovery in Databases,* KDD que nous emploierons par la suite pour décrire les méthodes de création de nouveaux savoirs à partir des bases de données bibliographiques. *Text-based Knowledge Discovery* est également un terme employé dans ce sens. Cette expression peut recouvrir beaucoup de techniques, de manière assez générique, comme le *text mining* ou la classification. Dans ce travail, le KDD est pris au sens littéral, c'est-à-dire comme une méthodologie de création de nouveaux savoirs à partir de bases de données bibliographiques.

Dans leurs conceptions et de part leur utilisation, les bases de données bibliographiques ne contiennent que des informations datées, qui, dans l'absolu, ne portent en elles aucune nouveauté. La présence même d'une référence bibliographique d'un article dans une base de données élimine toute nouveauté : l'information qu'il contient est rendue publique et est accessible à tous. Même si l'utilisateur de la base de données ne connaît pas par avance la quantité

d'informations qu'il aura en réponse à une requête, le résultat qu'il obtiendra est attendue – pour peu qu'il formule clairement sa requête. Par exemple, "Acne Vulgaris/drug therapy"[MAJR], requête Medline au format de PubMed, listera les articles qui parlent de manière centrale des traitements de l'acné. Une lecture rapide des titres des premières citations mettra en évidence des molécules telles que la lymécycline, l'adapalène, la tétracycline, le peroxyde de benzoyle ou encore l'isotrétinoine. Rien de bien nouveau pour le dermatologue qui suit la littérature de près. L'utilisation des opérateurs booléens ne permet pas de progresser vers la mise à jour de nouvelles connaissances. Ainsi, l'utilisation du ET met en évidence des éléments communs à deux thèmes (huile de poisson ET maladie de Raynaud) ou – au mieux ? - l'absence d'élément commun qui, dans le premier travail de Swanson, révèlera que son hypothèse n'est supportée par aucun lien direct et sera peut-être à l'origine d'une nouvelle découverte.

Medline contenait en janvier 2005 plus de 13 millions de citations. Considérant que le croisement des informations contenues dans deux articles différents peut donner naissance à une nouvelle hypothèse, Medline avec un nombre formidable de paires d'articles est susceptible de receler un nombre d'hypothèses latentes considérables, environ 88.000 milliards. Plus la production scientifique augmente en même temps que les scientifiques se spécialisent dans des domaines de plus en plus complexes, plus il y aura de connexions cachées. La méthode de KDD de Swanson s'accommode donc bien de cette explosion d'informations dont elle profite pleinement. Cela peut paraître paradoxale. En effet, d'une part, l'immense majorité des utilisateurs des sources d'informations se trouve confrontée à la difficulté de rechercher l'information pertinente, difficulté imputable à l'augmentation du volume d'informations, alors que d'autre part, la méthode de Swanson tire partie de cette profusion et laisse entrevoir à ceux qui l'utilisent de nouvelles opportunités de découvertes.

Le contexte de l'industrie pharmaceutique

La majorité des publications abordant le travail de Swanson est d'origine académique ou universitaire. Si l'on se base sur la littérature, l'industrie ne semble pas encore s'être réellement penchée sur le sujet. Or, il est aujourd'hui courant de trouver des articles sur le datamining ou le traitement de grands volumes de données par des méthodes bioinformatiques dans les revues de *drug discovery*. A notre connaissance, seuls Mack et Hehenberger ont abordé le sujet dans ce cadre [Mack, 2002]. Revenant sur le défi que représente la gestion de volume croissant d'informations, ils exposent quelques méthodes d'extraction d'informations et abordent, entre autres, l'exemple de la maladie de Raynaud et de l'huile de poisson. Mack et Hehenberger insistent sur le fait que les technologies aujourd'hui utilisées dans la découverte de nouveaux médicaments produisent de grandes quantités de données : le séquençage de génomes, le séquençage de protéines, les puces à ADN, les tests à haut débit (HTS, *high throughput screening*), etc…

Le processus de recherche et développement d'un médicament est long car il faut compter une dizaine d'années avant de pouvoir mettre une molécule sur le marché. Le taux d'attrition est très élevé, puisqu'une molécule sur 10.000 aura peut-être la chance d'être mise sur le marché [Lawrence, 2002]. Le coût de développement d'un médicament est estimé en moyenne à 802 millions de dollars US [DiMasi, 2003]. Les laboratoires pharmaceutiques prennent donc beaucoup de risques car ils ne peuvent compter que sur un nombre limité de produits pour financer leur R&D.

Aujourd'hui, l'industrie pharmaceutique a mis en place des stratégies de traitement à haut débit pour la phase précoce de découverte de molécules – *drug discovery* - [Warne, 2003] : il s'agit de gagner en productivité dans les étapes où il est possible d'automatiser les manipulations répétitives.

Nous verrons comment notre méthode peut intervenir comme outil d'aide à la sélection dans ces phases ou également être utilisée pour tenter de réorienter un médicament dans d'autres indications que celle pour laquelle il a été mis sur le marché.

Maladie de Raynaud et huile de poisson : la première découverte de Don Swanson et le modèle ABC

En 1986, Don Swanson, professeur à l'Université de Chicago, publie le premier article d'une longue série, dans lequel il expose une méthode originale de découverte de connaissances dans les bases de données bibliographiques [Swanson, 1986a]. Une des idées majeures qui a guidé les travaux de Swanson est qu'avec l'explosion du nombre de publications scientifiques et la fragmentation des communautés de chercheurs autours de thèmes toujours plus complexes, il existe certainement des connections latentes à découvrir. Supposons qu'un champ de la médecine lie une substance A avec des symptômes B et qu'un autre champ de la médecine, bien distinct du premier, lie ces mêmes symptômes B à une maladie C. Si ces deux faits sont décrits séparément dans la littérature, il existe une connexion cachée implicite et logique entre A et C, à travers B. Cependant, jusqu'à ce qu'un chercheur étudie de concert les littératures AB et BC, ce lien restera latent. Son travail sur la maladie de Raynaud[7] conduit Swanson à formuler l'hypothèse selon laquelle l'huile de poisson pourrait agir sur cette pathologie. A l'époque de ces travaux, il était bien établi que les patients atteints de la maladie de Raynaud avaient des problèmes d'agrégation plaquettaire et une viscosité sanguine élevée. Il était également connu que l'huile de poisson a pour effet, entre autres, d'inhiber l'agrégation plaquettaire et de diminuer la viscosité sanguine. Ces deux faits étaient largement repris à travers la littérature. Par contre, il n'existait aucune publication suggérant que l'huile de poisson pourrait traiter la maladie de Raynaud. Après analyse de la littérature, Swanson fut le premier à proposer d'utiliser l'huile de poisson comme traitement pour la maladie de Raynaud. On peut ainsi décrire le modèle de Swanson [Pierret, 2004] : entre un savoir sur une substance thérapeutique A et une maladie C, il existe des liens B, classiquement des phénomènes physiologiques. A travers la littérature biomédicale, les connaissances sur les liens AB et BC peuvent exister bien que la connexion

[7] La maladie de Raynaud est caractérisée par un arrêt temporaire de la circulation du sang au niveau des extrémités. Les doigts deviennent pâles et très douloureux. La maladie est favorisée par le froid.

implicite AC ne soit pas connue. Swanson a montré à plusieurs reprises [Swanson, 1986a, 1988, 1990a] que des parties disjointes de la connaissance biomédicale peuvent être connectées en étudiant leurs littératures respectives selon ce modèle transitif où :

- A désigne une substance active (bien souvent un médicament ou une substance chimique, mais également des vitamines, oligo-éléments, minéraux, protéines, …),
- B désigne les aspects physiologiques au sens large (physiopathologie) et l'anatomie,
- et C désigne les pathologies.

Plusieurs modes de transition sont possibles. Retenons les deux principaux :

- A→B→C, qui d'un point de départ A va permettre d'explorer différentes relations AB, puis BC. A l'origine, ni B ni C ne sont connus. Il s'agit d'un processus ouvert, utilisé pour générer une hypothèse (C→B→A est bien entendu inclus dans ce cas).
- A→B←C, qui va explorer les différentes relations B possibles entre A et C, processus clos permettant de tester une hypothèse.

Partant de la théorie de Swanson, nous avons mis au point une méthode de KDD, basée sur l'exploitation d'un thesaurus biomédical qui combine des séquences de traitements de la littérature à des analyses bibliographiques ciblées dans le but de générer ou de tester des hypothèses. Cette méthode a été appelée DPM (*Diseases Physiopathology Molecules*). Le lien entre une maladie et le médicament idéal pour la traiter est l'ensemble des phénomènes physiopathologiques qui caractérisent la maladie, sur lesquels le médicament va agir.

Pour résumer

Le mythe du "savant homme" du 19ème a laissé la place, avec le développement des techniques, à une compartimentation des spécialités et à un accroissement formidable de la littérature disponible. Swanson explique que cela n'est pas sans poser de problèmes, qui méritent que nous les examinions. Tout d'abord, la plupart des disciplines scientifiques sont certainement reliées à d'autres de manière logique. Ensuite, il existe bien plus de combinaisons possibles entre discipline scientifiques qu'il y a de disciplines. Enfin, le système de structuration de l'information dans les bases de données bibliographiques n'est pas organisé pour exploiter et valoriser les connexions, beaucoup nous échappent. Ainsi sont créées un grand nombre d'unités de littérature, indépendamment les unes des autres, sans tenir compte des relations logiques qui peuvent les lier et donner naissance à de nouveaux savoirs : il s'agit du savoir public caché [Swanson, 1986b]. La science répond à sa propre croissance par une augmentation de la spécialisation en négligeant les connexions.

Le modèle de Swanson tire partie de cette organisation fragmentée et de plus en plus cloisonnée de l'information scientifique et de la quantité colossale de publications disponibles : en combinant des informations bibliographiques existantes, bien que non reliées, on peut créer de nouvelles connaissances. Les points importants du modèle sont l'absence de lien entre les deux informations et leur complémentarité : cela conditionne l'existence d'une relation cachée.

Pour terminer cette introduction, ce travail a pour double objet :

- d'exposer une méthode de KDD, simple et rapide dans sa mise en œuvre pour analyser l'information bibliographique, qui s'appuie sur un traitement original de la littérature biomédicale, sur l'expertise humaine et sur l'exploitation ciblée des connaissances à notre disposition [Pierret, 2005], d'une part,
- et d'autre part, de démontrer la robustesse de la théorie de Swanson, non pas à partir de la multiplication d'expériences, mais à partir de la

multiplicité de méthodes qu'elle a inspiré et qui conduisent toutes aux mêmes conclusions.

Plan de la thèse

Dans un premier chapitre, nous proposons de situer rapidement le contexte historique et épistémologique du travail de Don Swanson. Puis nous détaillerons le processus de sa première découverte, jetant les bases du modèle ABC. Nous montrerons que son modèle et sa méthode ont été repris pas plusieurs équipes, qui ont travaillé sur les mêmes sujets que lui, en employant des techniques différentes, arrivant aux mêmes conclusions.

Le chapitre 2, après une présentation sommaire de la National Library of Medicine et des services qu'elle propose, détaillera le MeSH, thesaurus biomédical qu'elle a développé pour indexer, entre autre, les citations de Medline. Nous exposerons ensuite, comment à partir du MeSH et de Medline, nous avons développé un système de découverte de connaissances basé sur la littérature biomédicale, le DPM. Nous examinerons à travers trois expériences comment le DPM a été mis au point et terminerons sur les limites et possibilités de cet outil.

Le chapitre 3 ouvre des pistes pour les développements futurs du DPM en proposant de s'affranchir du MeSH ou de travailler sur d'autres sources de données que Medline.

[…] par hasard diriez-vous peut-être, mais souvenez-vous que, dans les sciences d'observation le hasard ne favorise que des esprits préparés […]

Louis Pasteur

PREMIERE PARTIE

Etat de l'art

Afin de simplifier la lecture de la bibliographie de ce manuscrit, nous l'avons séparé en deux parties. La bibliographie générale, dont les références figurent en police normale [Swanson, 1986b], rapportée dans la section bibliographie. La bibliographie sur laquelle Don Swanson s'est appuyé pour ses travaux, dont les références figurent en italique *[Dyerberg, 1982]* et sont présentées dans l'annexe 1.

1.1 Historique de la découverte de Don Swanson

Don R. Swanson est physicien de formation et a manifesté tout au long de sa carrière un grand intérêt pour l'information biomédicale. Professeur émérite de l'Université de Chicago, il a reçu la plus haute distinction de l'ASIST[8] en 2000 (ASIST *Award of Merit*) pour l'ensemble de ses travaux sur le KDD.

Au début des années 80, Don Swanson remarque un article sur l'alimentation des esquimaux. La consommation de poissons et de mammifères marins, riches en acides gras poly-insaturés longs, diminue le facteur de risque de maladies cardiovasculaires, d'où leur moindre incidence chez les esquimaux *[Dyerberg,*

[8] American Society for Information Science and Technology. (Page consultée le 22 septembre 2005). *American Society for Information Science and Technology*, [En ligne]. Adresse URL : http://www.asis.org

1982] et [Dewailly, 2001]. Swanson effectue alors une série de recherches bibliographiques dans ce sens et il trouve que :

- l'huile de poisson, composée en grande partie de tels acides gras, était connue pour diminuer la viscosité du sang et l'agrégation des plaquettes (favorise la prévention des thromboses et de l'athérosclérose) et pour agir sur la réactivité vasculaire, d'une part,
- et d'autre part, dans la maladie de Raynaud la viscosité du sang et l'agrégation plaquettaire augmentent et il se produit une vasoconstriction exagérée.

Le lien est évident et Swanson fut le premier à formuler l'hypothèse selon laquelle l'huile de poisson est un traitement potentiel de la maladie de Raynaud. En effet, avant 1986, aucun document ne lie l'huile de poisson et la maladie de Raynaud. Une publication détaille son hypothèse d'un point de vue physiologique [Swanson, 1986a] et une autre expose brièvement la méthode employée [Swanson 1987]. En 1989, une équipe de cliniciens d'Albany Medical College à New York montre que même si l'huile de poisson ne permet pas de guérir de la maladie de Raynaud, elle contribue à améliorer l'état des malades [DiGiacomo, 1989].

Swanson résume ainsi le contexte de sa découverte : "*In 1985, I was struck by lightning and have never recovered*" [Swanson, 2001a]. Il a réalisé que deux informations issues d'articles médicaux différents suggèrent, lorsqu'on les juxtapose, une hypothèse que personne ne connaissait alors. La connexion de deux informations disjointes peut créer une nouvelle connaissance. Son approche était plus intuitive que structurée. En 1986, dans un article publié un an plus tard, il regrette de ne pouvoir décrire de processus systématique de recherche de connexions cachées [Swanson, 1987]. Mais il élabore rapidement une stratégie basée sur l'utilisation de bases des données bibliographiques Medline, Embase et SciSearch, baptisée *explore/exclude* ou *trial-and-error*. Cette stratégie permet de rechercher les connections entre deux articles (*literatures*), non interactifs (ne se citent pas) et complémentaires afin de générer une nouvelle information absente des deux articles considérés

séparément [Swanson, 1989a]. Son travail portera principalement sur l'amélioration de sa méthode de KDD et la découverte de nouvelles hypothèses.

1.2 Le cadre épistémologique

Bien avant de publier ses travaux sur la maladie de Raynaud et l'huile de poisson, Don Swanson s'est intéressé à la diffusion de l'information biomédicale, principalement dans l'optique d'en améliorer l'accès. Il proposait de travailler, entre autre, sur la précision et le rappel par l'utilisation d'index de citations d'articles et de la notion de couplage bibliographique [Swanson, 1974]. Swanson décrira la recherche d'information, dans le cadre d'un travail scientifique, comme un processus *trial-and-error* [Swanson, 1977]. Selon lui, la recherche d'information est un processus proche de celui qui conduit à élaborer une théorie scientifique, c'est-à-dire le travail de recherche scientifique. La découverte scientifique ne commence pas avec un sujet, mais avec un problème, le chercheur étant quelqu'un de curieux, persévérant, préoccupé par ce problème. Une hypothèse initiale ou au moins un embryon de solution doit également préexister dans l'esprit du chercheur. La théorie ne naît pas de l'observation. De nouvelles observations peuvent conduire à une théorie, seulement en corrigeant ou en modifiant une théorie préexistante. La connaissance grandit par un processus d'essais et d'erreurs. De manière similaire, la recherche d'information a pour base une hypothèse ou une conjecture et est guidée par une idée que le chercheur souhaite tester. L'attrait principal de la technique *trial-and-error* ne réside pas tant dans son usage direct pour retrouver des documents pertinents que dans le fait qu'elle permet de reformuler une requête. La requête, point de départ de la recherche d'information, est la description imparfaite par le chercheur des attributs qu'il considère qu'un document relavant doit posséder. C'est une estimation – une conjecture - qu'il teste en examinant les documents ramenés par cette requête. Chaque article ainsi trouvé doit être principalement considéré comme un stimulus pour un nouvel essai de requête, sans avoir de critères absolus permettant de dire que la recherche d'information est terminée. Parce que

chaque article contribue généralement à dessiner une nouvelle requête, le chercheur apprend en corrigeant ses erreurs. Retrouver des documents non relevants peut parfois s'avérer important dans l'amélioration de la requête, tout comme le font les documents pertinents. Se tromper, faire des erreurs est essentiel dans le processus de recherche. Swanson est convaincu que le processus *trial-and-error* joue un rôle central dans la recherche documentaire. Il est intéressant de noter que, dans ce papier, Swanson discute ensuite du concept de pertinence – *relevance* – non dans le but d'en donner une définition absolue, mais plutôt pour aider à dessiner ou évaluer des systèmes de recherche documentaire. Il nous livre deux propositions. Sa première est – du point de vue du présent travail – la plus intéressante, puisqu'un document est défini comme pertinent s'il peut être pris comme un élément qui permet au chercheur de créer un nouveau savoir, au regard de son besoin d'information. Ce nouveau savoir est très subjectif et ne peut être estimé que par le chercheur lui-même. Dans ce sens, la pertinence ne peut pas vraiment être mesurée et c'est un critère qui ne peut être assigné aux documents que par le chercheur lui-même. En quelque sorte, selon cette proposition, la *relevance* est une supposition inscrite dans le cadre du processus *trial-and-error*, proposition pour laquelle le document jugé pertinent a servi de stimulus. Sa seconde proposition, dit que la *relevance* est synonyme de "du même sujet", un article pertinent, dans ce contexte traite du sujet à partir duquel – ou pour lequel – la requête a été formulée. Ici, la pertinence est synonyme de trivialité puisque les documents sélectionnés par une requête sont attendus.

Deux ans plus tard, Swanson livre sa réflexion sur la place des bibliothèques dédiées à la recherche face à la croissance de la connaissance, exposant de ce point de vue les difficultés que rencontrent les chercheurs et savants dans leurs tentatives de construire de nouvelles connaissances [Swanson, 1979]. Pour Swanson, la croissance des connaissances est une question centrale qui doit être pris en compte dans le développement et l'organisation des bibliothèques : il suggère que l'accès à l'information soit pensé de manière à favoriser au maximum la croissance de la connaissance. Parallèlement à un processus d'amélioration perpétuel basé sur une analyse critique du fonctionnement de ces

bibliothèques, Swanson propose que l'accès à la littérature se fasse par problématique et non pas par sujet. Un accès "orienté problème" faciliterait le travail des chercheurs qui tentent de lier des informations disjointes dans le but de créer un savoir cohérent. Il relève au passage que le travail d'Eugène Garfield, avec le Science Citation Index, va dans ce sens, puisqu'il permet au chercheur de connaître le devenir d'un article et de suivre ses citations dans les différentes publications qui s'y réfèrent.

Dans ses travaux, Don Swanson fait de nombreuses références à l'épistémologie de Karl Popper [Swanson, 1977, 1979, 1986b, 1993]. Le processus *trial-and-error* renvoie à la connaissance objective : en deux mots, une théorie est construite par élimination des erreurs contenues dans les théories précédentes. Le modèle des 3 mondes de Popper trouve également un écho particulier dans les travaux de Swanson [Bawden, 2002]. Les 3 mondes offrent un cadre à l'étude de l'information et du savoir :

- monde 1 : monde physique des objets.
- monde 2 : monde de la subjectivité individuelle, monde mental.
- monde 3 : monde des idées et de la culture humaine dans son ensemble.

Bien que critiquée [Robillard, 2004], la théorie des 3 mondes sert de support de réflexion dans le domaine des sciences de l'information. En particulier, David Bawden insiste sur son utilité dans le cas de l'information et du savoir relatif à la santé et aux soins [Bawden, 2002]. Ainsi, il propose :

- monde 1 : les personnes, ouvrages, ordinateurs.
- monde 2 : le monde intérieur des êtres pensants.
- monde 3 : le savoir objectif, communicable, incluant les ouvrages, les bases de données, les bibliothèques ou tout autres supports d'information, non pas pour leurs formes, mais pour leurs contenus (au passage, une des critiques souvent formulée à l'encontre de Popper est que le monde 3 contient des éléments du monde 1).

Le monde 3 de Popper est un concept central pour Swanson, qui permet d'appréhender la création de savoir. Le monde 3 a une existence indépendante et autonome où le "savoir objectif" préexiste à l'activité humaine (monde 2) qui a conduit à sa découverte. Le chercheur accède au monde 3 par une méthode d'élimination des erreurs contenues dans les théories sur lesquelles il travaille. La stratégie *trial-and-error* de Swanson s'inscrit bien dans ce modèle, l'utilisant comme une base pour explorer le monde abstrait de la connaissance, connaissance de plus en plus fragmentée par la spécialisation des disciplines scientifiques. Il est ainsi possible de faire de nouvelles découvertes en exploitant les liens cachés entre diverses parties du savoir biomédical.

1.3 Maladie de Raynaud et huile de poisson

La première découverte de Swanson est en partie fortuite. En partie seulement, car par intérêt personnel, Don Swanson lit beaucoup d'articles médicaux, en particulier sur la maladie de Raynaud et la migraine. 1986 sera l'année charnière pour la mise au point de son modèle de KDD, année de sa première découverte. Il écrira deux articles dans lesquels il expose sa théorie sur le lien entre l'huile de poisson et la maladie de Raynaud. Le premier [Swanson, 1986b], publié en avril, amorce en quelque sorte la transition entre ses articles des années 70, portant sur les théories de l'accès et de la diffusion de l'informations et les articles plus récents, plus appliqués, exposant soit ses hypothèses, soit les techniques développées autour de son modèle. Le second [Swanson, 1986a], le plus important à nos yeux, présente de manière extensive les liens qu'il a découvert entre la maladie de Raynaud et l'huile de poisson, à l'aide d'une bibliographie rigoureusement sélectionnée.

1.3.1 Introduction du modèle ABC

Poursuivant sa réflexion sur les 3 mondes de Popper [Swanson 1986b], Swanson illustre comment exploiter le monde 3 pour révéler des savoirs latents. Le processus de recherche d'information, à l'instar d'une théorie scientifique, peut être critiqué et amélioré, mais on ne pourra jamais vérifier qu'il est capable de

retrouver toute l'information pertinente pour un problème ou une théorie donné, l'exhaustivité n'étant pas assurée. C'est un processus imparfait ce qui, aux yeux de Swanson, le rend particulièrement intéressant dans le contexte de la découverte de connaissances. C'est sur cette faiblesse que repose la notion du savoir public non découvert (*undiscovered public knowledge*). Faiblesse à double tranchant, qu'il convient d'exploiter dans le bon sens. Ainsi, la proposition "tous les chats que nous avons rencontré étaient gris, donc tous les chats sont gris" est vraie tant que nous ne rencontrons pas un chat d'une autre couleur. L'observation d'une multitude de chats gris, n'assure pas la véracité de cette proposition alors que l'observation d'un seul chat noir l'infirme. C'est toute la fragilité du raisonnement inductif basé sur l'exploitation de données incomplètes, non exhaustives. D'un autre coté, la fragmentation des connaissances qui ajoute à l'impossibilité d'effectuer une recherche complète d'information, ouvre le champ à la découverte de nouveaux savoirs, basée sur la complémentarité d'éléments disponibles dans la bibliographie. Ainsi, si A cause B et B cause C, alors A cause ou implique C. Si les relations AB et BC sont connues, mais ne sont liés d'aucune manière, alors AC restera caché. Pour que ce lien soit découvert, il faut qu'une même personne au même moment s'intéresse à AB et BC et perçoive l'implication logique AC. Remarquer que A implique C est la découverte d'une nouvelle connaissance par un explorateur du monde 3. Swanson suggère que la recherche d'information et le savoir public non découvert, bien qu'illimités en diversité et complexité, présentent une certaine structure, un certain ordre qui mérite d'être étudié plus en avant. La croissance du savoir scientifique est habituellement perçue comme porteuse de nouvelles découvertes à partir du monde 1. Cependant, les professionnels de l'information pourraient explorer et exploiter le monde 3 afin d'en dégager de nouvelles connaissances. Swanson n'oppose pas la découverte de connaissances par l'exploitation de la littérature à la découverte de connaissances par l'expérimentation. Son propos est d'examiner la structure logique des connexions cachées et le processus de découverte de telles connexions. Les théories scientifiques dépendent étroitement des connaissances disponibles dans

la littérature et il est illusoire de vouloir séparer les découvertes basées sur la littérature de celles basées sur l'expérimentation au laboratoire.

1.3.2 Méthode bibliographique

Swanson a formulé sa première hypothèse par l'analyse de 59 articles répartis en deux groupes [Swanson, 1986a] :

- 34 articles sur la maladie de Raynaud.
- 25 articles sur l'huile de poisson.

Ces deux "littératures" sont parfaitement disjointes : en novembre 1985, une recherche bibliographique conduite sur Embase et Medline montre qu'il n'existe aucun article commun entre la bibliographie relative à la maladie de Raynaud et celle relative à l'huile de poisson. Ces concepts sont pris au sens large et incluent :

- maladie de Raynaud et sclérodermie (maladie souvent accompagnée du syndrome de Raynaud),
- huile de poisson, acide eicosapentaenoic (EPA), huile de foie de morue, huile de saumon, huile de menhaden et esquimaux.

Qui plus est, aucun des articles du groupe huile de poisson ne cite un seul article du groupe maladie de Raynaud et réciproquement. Ces deux groupes sont mutuellement isolés. En revanche, au sein de chaque groupe les liens sont forts [Swanson, 1987] :

- huile de poisson : 18 articles citent 34 fois 12 articles.
- maladie de Raynaud : 27 articles citent 64 fois 18 articles.

La quantité d'articles relatifs à la maladie de Raynaud publiés entre 1975 et 1985 est de l'ordre de 2.000 et dans le même temps, on compte environ un millier d'articles sur l'huile de poisson. Les 59 articles sélectionnés par Swanson représentent une petite part de l'information disponible à l'époque sans la moindre signification statistique. Ils ne servent qu'à soutenir l'argument logique qui lie l'huile de poisson et la maladie de Raynaud.

1.3.2.1 Etude des co-citations

L'étude des co-citations part de l'idée que des relations peuvent exister entre documents qui sont cités ensemble. Par exemple, deux documents qui ne sont cités qu'une seule fois ensemble n'auront peut-être pas de relation entre eux. En revanche, deux documents cités ensemble à de nombreuses reprises par différents autres documents seront probablement liés. Swanson a réalisé l'étude des co-citations pour estimer à quel point les deux littératures sont disjointes. Son but est de rechercher un ou des articles quelconques qui citent en même temps un article de chaque littérature au moins.

L'analyse des citations d'un ou plusieurs de ces 59 articles à l'aide de SciSearch identifie 489 articles citants [Swanson, 1986a] et [Swanson, 1987]. 173 de ces articles citent plus d'une fois un élément des littératures sur l'huile de poisson ou la maladie de Raynaud, ce qui permet d'isoler 194 paires d'articles différentes, co-citées 500 fois. Seulement quatre sur 489 citent à la fois au moins un article parmi les 34 sur la maladie de Raynaud et au moins un article parmi les 25 sur l'huile de poisson. Ces quatre articles sont des revues traitant des prostaglandines. Deux ne présentent aucun intérêt dans notre contexte, puisque la maladie de Raynaud n'est citée qu'une seule fois par article, indépendamment de l'huile de poisson *[Preston, 1985]* et *[Weksler, 1984]*. De plus, leurs vastes bibliographies contribuent à diluer l'impact de la présence simultanée des articles des deux groupes (respectivement 66 *[Preston, 1985]* et 113 citations *[Weksler, 1984]*). Les deux autres articles sont pratiquement identiques en terme de contenu et d'auteur ; Swanson considère qu'il s'agit d'une même revue de la littérature *[Moncada, 1983, 1984]*. Bien qu'ils ne mentionnent pas de lien entre l'huile de poisson et la maladie de Raynaud, le fait qu'ils exposent les deux sujets – de manière séparée – suggère un rôle important de ce type d'article. Dans ce cas, les deux sujets sont discutés isolément, mais sur des bases physiologiques communes. Les revues de la littérature, de portée générale, qui proposent un aperçu global d'un sujet, offre un support au rapprochement de deux sujets disjoints mais complémentaires (sans que les auteurs n'en aient forcément conscience). Pour ces deux articles, l'importance de la bibliographie citée tend à cacher la connexion (198 pour celui de 1983 et 159 pour celui de

1984). Enfin, Swanson a regardé un par un les titres des 489 articles, identifiant deux autres articles intéressants. L'un suggère de traiter le syndrome de Sjorgren et la sclérodermie avec des acides gras essentiels *[Horrobin, 1984]*. L'autre décrit une étude clinique de l'huile essentielle de primevère (EPO, *evening primerose oil*) comme traitement de la maladie de Raynaud *[Belch, 1985b]*. Ces quatre articles identifiés par Swanson constituent la seule relation, ténue, entre la maladie de Raynaud et l'huile de poisson. Leur analyse ne permet pas d'envisager ce lien.

Bien que les deux littératures fassent l'objet d'un grand nombre de co-citations, très peu de ces co-citations concernent à la fois l'huile de poisson et la maladie de Raynaud. De ce point de vue, les deux littératures sont isolées. Le tableau 1 résume l'état des co-citations dans ces quatre articles.

Articles citants	Articles cités		Intérêt
	Groupe maladie de Raynaud	**Groupe huile de poisson**	
Prostacyclin and its clinical applications *[Moncada., 1984]*	Intermittent epoprostenol (prostacyclin) infusion in patients with Raynaud's syndrome. A double-blind controlled trial *[Belch, 1983]* Treatment of Raynaud's phenomenon by intravenous infusion of prostacyclin (PGI2) *[Dowd, 1982]*	The long-term effect of dietary supplementation with fish lipid concentrate on serum lipids, bleeding time, platelets and angina *[Saynor, 1984]* Beneficial effect of fish oil on blood viscosity in peripheral vascular disease *[Woodcock, 1984]*	+
Biology and therapeutic potential of prostacyclin *[Moncada., 1983]*	Treatment of Raynaud's phenomenon by intravenous infusion of prostacyclin (PGI2) *[Dowd, 1982]*	The effect of cod liver oil and corn oil on platelets and vessel wall in man *[Brox, 1981]*	+
Platelet suppressive therapy in clinical medicine *[Preston, 1985]*	Intermittent epoprostenol (prostacyclin) infusion in patients with Raynaud's syndrome. A double-blind controlled trial *[Belch, 1983]*	Beneficial effect of fish oil on blood viscosity in peripheral vascular disease *[Woodcock, 1984]*	–
Prostaglandins and vascular function *[Weksler, 1984]*	Intermittent epoprostenol (prostacyclin) infusion in patients with Raynaud's syndrome. A double-blind controlled trial *[Belch, 1983]*	Vascular reactivity and high dietary eicosapentaenoic acid *[Lockette, 1982]*	–

Tableau 1 : Bilan des co-citations communes aux littératures sur la maladie de Raynaud et sur l'huile de poisson

1.3.2.2 Etude du couplage bibliographique

Swanson a également étudié les articles cités par ses deux littératures [Swanson, 1987], pour appréhender leur degré de couplage toujours dans l'idée d'estimer la séparation qui existe entre elles. A l'opposé de l'étude des co-citations, l'étude du couplage cherche à savoir si deux articles donnés citent un même article. Ainsi, 98 articles cités couplent les 25 articles sur l'huile de poisson et 100 couplent les 34 sur la maladie de Raynaud. En étudiant le nombre de citations communes à deux articles, Swanson remarque que le couplage est plus fort entre les éléments du groupe de l'huile de poisson. 91% des articles sur l'huile de

poisson sont couplés (dont 41% par quatre articles communs ou plus) contre 30% des articles sur la maladie de Raynaud.

Le couplage entre les deux groupes d'article existe, puisque 13 paires sont couplées. Sont impliqués 7 articles sur la maladie de Raynaud et 10 articles sur l'huile de poisson. L'analyse détaillée de ces paires d'articles montre que :

- 6 le sont par la citation d'articles sur des techniques de laboratoire, qui n'apportent pas d'élément logique au lien entre l'huile de poisson et la maladie de Raynaud.
- 2 citent deux aspects complètement différents du même article.
- 3 sont couplés parce qu'ils se réfèrent au même savoir général, porteur de peu ou aucune signification dans notre contexte.
- enfin, les 3 dernières impliquent deux articles du groupe sur l'huile de poisson, articles semblables sur la thématique et les auteurs - *[Cartwright, 1985]* et *[Woodcock, 1984]* - et un article sur la maladie de Raynaud *[Blunt, 1980]*. Ces trois articles citent la même référence *[Dormandy, 1973]* pour la même raison : il traite de l'hyperviscosité sanguine. Cet élément est pertinent en regard de la physiopathologie de la maladie de Raynaud.

Ne considérant que les paires d'articles composées d'un élément de chaque littérature, sur les 975 combinaisons possibles, il n'en existe que 13 dont une seule est pertinente. En novembre 1985, les deux savoirs étaient assurément disjoints, leurs littératures respectives étant presque complètement non interactives.

Le tableau 2 illustre l'étude du couplage bibliographique au sein de chaque littérature :

		Maladie de Raynaud	Huile de poisson
Nombre d'articles		34	25
Nombre de paires d'articles théorique		561	300
Nombre de paires couplées		170 (30%)	273 (91%)
Nombre de paires non couplées		391 (70%)	27 (9%)
Nombre d'articles cités en commun (force du couplage)	1	97	48
	2	32	61
	3	26	51
	≥4	15	113

Tableau 2 : Bilan de l'analyse du couplage bibliographique au sein de chacune des deux littératures étudiées par Swanson

1.3.2.3 Analyse des littératures complémentaires : effet plausibles de l'huile de poisson sur la maladie de Raynaud

L'article de Dyerberg *[Dyerberg, 1982]* semble être à l'origine de la réflexion de Swanson. L'auteur y avance l'hypothèse que l'huile de poisson, en particulier l'acide eicosapentaenoique (EPA) pourrait prévenir les thromboses et l'athérosclérose. Cette hypothèse est basée sur l'étude d'esquimaux du Groenland chez qui l'incidence de l'infarctus – conséquence, entre autres de l'athérosclérose ou de thromboses – est très faible. Ces populations ont des taux sanguins de cholestérol et triglycérides très bas. Cela est attribué à leur régime alimentaire particulièrement riche en EPA trouvé chez les animaux marins qui vivent en eau froide.

Swanson utilise les 25 articles de la littérature sur l'huile de poisson pour montrer que :

- l'huile de poisson, essentiellement l'EPA, réduit l'agrégation plaquettaire, inhibe la vasoconstriction, diminue la viscosité sanguine et favorise la déformation des érythrocytes (globules rouges, qui s'ils sont trop rigides augmentent la viscosité du sang),
- l'EPA est transformé en prostaglandine I3 (PGI3), de structure proche de la PGI2 (ou prostacycline) et ayant les même activités vasodilatatrice et antiagrégante,
- l'EPA ou l'huile de poisson réduisent les niveaux lipidiques sanguins, notamment des triglycérides dont le taux est directement corrélé à la viscosité du sang *[Ozanne, 1984]* et *[Seplowitz, 1981]*.

En utilisant les 34 articles de la littérature sur le maladie de Raynaud, il met en évidence que :

- la maladie de Raynaud est caractérisée par une augmentation de l'agrégation plaquettaire (activation plaquettaire), une rigidité des érythrocytes, une augmentation de la viscosité sanguine – en particulier en cas de syndrome de Raynaud associé à d'autres pathologies comme la sclérodermie – et par une vasoconstriction,
- la prostaglandine E1 (PGE1), la PGE2 ou la nifedipine utilisées pour traiter la maladie de Raynaud ont des propriété antiagrégantes et vasodilatatrices,
- la kétansérine, diminuant la viscosité sanguine a également été utilisée pour traiter la maladie de Raynaud, même si des résultats contradictoires ont été obtenus.

A partir de ces éléments, en se plaçant dans le contexte bibliographique de 1985, il est logique de proposer que l'huile de poisson pourrait être un traitement de la maladie de Raynaud. Les bases de l'argumentaire sont solides. Le tableau 3 reprend très succinctement la complémentarité des deux littératures.

	Maladie de Raynaud	Huile de poisson
Agrégation plaquettaire	Platelet activation, fibrinolytic activity and circulating immune complexes in Raynaud's phenomenon *[Kallenberg, 1982]*	Inhibition of platelet aggregation and thromboxane synthesis after intake of small amount of icosapentaenoic acid *[Driss, 1984]*
Vasoconstriction	Calcium entry blocking agents in digital vasospasm (Raynaud's phenomenon) *[Kahan, 1983]*	Vascular reactivity and high dietary eicosapentaenoic acid *[Lockette, 1982]*
Viscosité sanguine	Hyperviscosity and thrombotic changes in idiopathic and secondary Raynaud's syndrome *[Blunt, 1980]*	Beneficial effect of fish oil on blood viscosity in peripheral vascular disease *[Woodcock, 1984]*
Déformabilité érythrocytaire	Increased prostacyclin metabolites and decreased red cell deformability in patients with systemic sclerosis and Raynauds syndrome *[Belch, 1985a]*	Effect of oral administration of highly purified eicosapentaenoic acid on platelet function, blood viscosity and red cell deformability in healthy human subjects *[Terano, 1983]*

Tableau 3 : Complémentarité des littératures sur l'huile de poisson et sur la maladie de Raynaud

La critique que l'on peut adresser à cette analyse est que l'étude de l'isolation des deux littératures est basée sur seulement 59 articles. C'est à partir de ce petit nombre que Swanson explore les citations et le couplage. Peut-être existait-il des liens similaires entre les deux littératures, mais n'impliquant pas les articles sélectionnés par Swanson ? Cependant, il est bien réel qu'au premier degré, huile de poisson et maladie de Raynaud étaient bien disjoints, c'est-à-dire qu'aucun article ne traitait des deux sujets.

1.4 Migraine et magnésium, une seconde découverte à partir de la méthode bibliographique

Reprenant le même mode d'exploitation des bases de données bibliographiques que pour l'exemple sur la maladie de Raynaud, Swanson publie une seconde série d'articles où il explore les liens entre la migraine et le magnésium [Swanson, 1988, 1989b, 1990b]. Ses travaux sur Medline et SciSearch datent d'août 1987. A cette époque, seulement 6 articles mentionnent migraine et magnésium contre 4.600 sur la migraine et 38.000 sur le magnésium. Un nombre aussi petit peut suggérer que les deux littératures sont pratiquement non interactives. Il est donc plausible qu'il existe entre elles des connexions logiques inconnues. L'étude de la littérature révèle que quelques auteurs ont effectivement abordé la question sans détailler les liens entre migraine et magnésium. Il n'existe pas d'étude contrôlée sur l'usage du magnésium dans le traitement de la migraine. Les co-citations confirment que les deux littératures sont isolées. Swanson analyse, décortique 65 articles sur la migraine et 63 sur le magnésium. Il en dégage 11 connexions logiques, cachées qui lient migraine et magnésium. Sans entrer dans le détail de l'analyse bibliographique, le tableau 4 expose ces connexions.

	Arguments de la littérature sur la migraine	Arguments de la littérature sur le magnésium
1	Stress et personnalité de type A[9] sont associés à la migraine	Stress et personnalité de type A conduisent à une perte globale de magnésium
2	Réactivité et tonus vasculaires excessifs peuvent augmenter la susceptibilité à la migraine	Le magnésium peut réduire le tonus et la réactivité vasculaire
3	Les bloqueurs de canaux calciques[10] peuvent prévenir la migraine	Le magnésium est un bloqueur naturel des canaux calciques
4	La dépression corticale extensive[11] est probablement impliquée dans la phase précoce de la migraine	Des taux élevés de magnésium extracellulaire cérébral peuvent inhiber la dépression corticale extensive
5	Migraine et épilepsie sont liées	Une déficience en magnésium est un facteur favorisant l'épilepsie
6	Les patients souffrant de migraine ont une agrégabilité plaquettaire élevée	Le magnésium inhibe l'agrégation plaquettaire
7	Chez les patients atteints de migraine, les plaquettes libèrent de la sérotonine de manière anormale	Le magnésium peut inhiber les contractions des vaisseaux sanguins induites par la sérotonine
8	La substance P peut être à l'origine de la douleur ressentie dans la migraine	Le magnésium peut supprimer l'activité de la substance P
9	Une libération anormale de prostaglandines peut aggraver la vasoactivité pendant la migraine	Le magnésium augmente la formation de prostacycline (PGI2), vasodilatatrice
10	La migraine peut impliquer un processus inflammatoire	Le magnésium a des propriétés anti-inflammatoires
11	L'hypoxie cérébrale peut jouer un rôle clé dans la migraine	Le magnésium peut protéger le cerveau contre les dommages engendrés par une hypoxie

Tableau 4 : Migraine et magnésium, 11 connexions logiques

A travers ces 11 arguments, Swanson montre que le magnésium semble être logiquement impliqué dans chaque étape de l'enchaînement complexe

[9] Comportement caratérisé par une conduite excessive et ambitieuse, de l'impatience, un sens profond de la compétition et de l'urgence et une agressivité peu contenue.

[10] Classe de médicaments qui induit la relaxation des muscles lisses. Ils sont utilisés dans le traitement de l'hypertension, des spasmes cérébrovasculaires, comme agents protecteurs du myocarde et comme relaxant dans le cas de spasmes utérins.

[11] La dépression corticale extensive est une onde de dépolarisation des neurones et des cellules gliales impliquées dans des troubles de la régulation neurovasculaire comme les accidents vasculaires cérébraux, les traumatismes craniens ou la migraine.

d'évènements conduisant à un épisode de migraine. Ces éléments confortent de manière implicite mais non révélée les théories jusqu'alors avancées pour expliquer le mécanisme de la migraine. Il est particulièrement frappant de noter que la migraine comprend des composantes à la fois vasculaires et neurologiques et que le magnésium joue sur la réactivité vasculaire et neurologique.

On relève une centaine d'articles abordant le sujet dans Medline après 1987. L'importance du magnésium est clairement établie, bien que son rôle précis dans le développement de la migraine reste à découvrir [Mauskop, 1998]. Trois études cliniques (dont deux randomisées en double-aveugle) montrent qu'il y a un bénéfice pour le patient atteint de migraine à prendre du magnésium [Pfaffenrath, 1996], [Peikert, 1996] et [Mauskop, 1995]. Ces trois études citent l'article original de Swanson de 1988.

Swanson et Smalheiser étendront cette analyse quelques années plus tard [Smalheiser, 1994]. Se basant sur le rôle que jouent les récepteurs NMDA[12] dans la pathogenèse de maladies du système nerveux central et sur le lien entre ce récepteur et le magnésium, ils proposent d'étudier comment le magnésium alimentaire peut agir sur l'apparition de ces maladies. Une déficience chronique en magnésium induirait une excitotoxicité des récepteurs NMDA et par conséquent un risque plus élevé de développer des maladies telles que des neurodégénérations ou liées au tonus vasculaire cérébral.

1.5 La méthodologie *explore/exclude* ou *trial-and-error*

Les deux premières hypothèses formulées par Don Swanson, à partir d'une exploitation rigoureuse de la littérature ont ensuite, avec plus ou moins de succès, eut un écho auprès des cliniciens et chercheurs. Certes, l'utilisation de l'huile de poisson dans la maladie de Raynaud n'a fait l'objet que d'un seul essai clinique concluant [DiGiacomo, 1989] et depuis 1986 on relève seulement

[12] Récepteurs N-Methyl-D-Aspartate, impliqués dans la plasticité neuronale et la toxicité due à une excitation neurnonale (excitotoxicité).

une quinzaine d'articles sur ce sujet[13]. En revanche, la littérature sur l'utilisation du magnésium dans la migraine est plus conséquente puisque depuis 1988, 150 articles ont été publiés[12].

Nous proposons d'analyser plus en avant la méthode bibliographique employée pour réaliser ces deux découvertes.

S'appuyant sur l'exemple de la maladie de Raynaud et de l'huile de poisson, Swanson proposera une méthodologie de KDD [Swanson, 1989a]. Elle se décompose en deux parties de deux étapes. Il s'agit d'abord d'analyser la littérature sur un sujet donné, pour identifier les connections logiques qui caractérisent ce sujet. C'est l'étape exploratoire qui fait appel à la créativité humaine et la stimule pour identifier des ensembles d'articles logiquement reliés. Puis la seconde partie a pour objectif d'exclure toutes les connections connues et interactives. Medline est la base de données bibliographiques utilisée.

1.5.1 Première partie : exploration

Etape 1 : dans la littérature biomédicale, les titres des articles signalent souvent des concepts en relation avec le thème principal de l'article. Swanson exploite cette particularité pour rechercher les facteurs sur lesquels agir afin de traiter ou soulager les personnes atteintes de la maladie de Raynaud. Il identifie ainsi la viscosité du sang et d'autres propriétés hémorhéologiques[14]. Les articles doivent être sélectionnés selon des critères qui permettent d'éclairer les relations possibles. Ici, deux critères sont retenus. Les titres des articles doivent contenir le terme "*Raynaud*", d'une part et, d'autre part, les articles sélectionnés doivent également contenir "*Raynaud Disease*" comme descripteur afin de ne travailler que sur certains qualificatifs. Medline est indexée avec le

[13] D'après Embase et Medline à l'exlusion des papiers reliés aux travaux de Don Swanson.

[14] Hémorhéologie : étude de la circulation sanguine en relation avec la pression, le flux sanguin, le volume, la résistance des vaisseaux sanguin et la viscosité.

MeSH, qui contient à la fois des descripteurs et des qualificatifs. Les qualificatifs donnent la possibilité de préciser le contexte dans lequel le descripteur est employé. Pour cette étape, Swanson propose de regarder, par exemple, "*Raynaud Disease/Blood*", pour ne sélectionner que les articles qui abordent les changements sanguins au cours de la maladie. Ce qui peut se traduire par la requête PubMed suivante :

```
"Raynaud Disease/blood"[MeSH:NoExp] AND raynaud*[TI]
```

En limitant la date à novembre 1985 au plus tard, cette requête rapporte 58 articles.

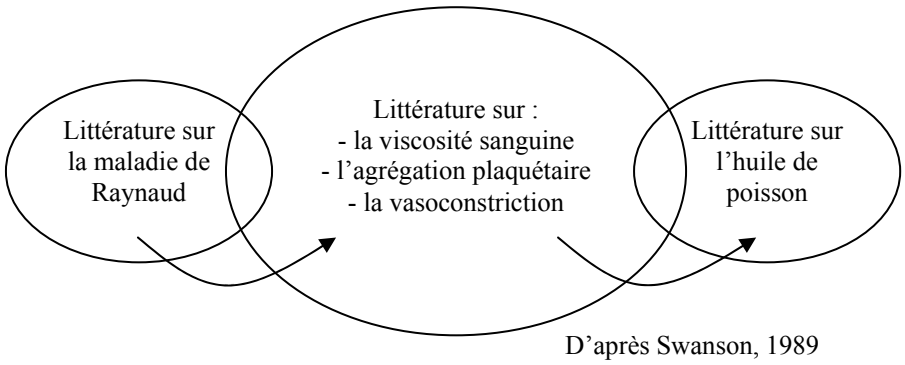
Etape 2 : une seconde recherche Medline sur la viscosité du sang identifie les moyens de la modifier, d'agir sur elle. C'est le but de cette seconde étape. Swanson propose une requête qui combine les concepts de viscosité et déformabilité des érythrocytes, termes devant apparaître dans le titre des articles. Nous pensons que sa requête devait être [15]:

```
visco*[TI] AND deformab*[TI]
```

Dans les conditions originales (i.e. novembre 1985), Swanson trouve 32 articles. Deux parlent directement de l'effet de l'huile de poisson sur ces deux paramètres sanguins *[Cartwright, 1985]* et *[Terano, 1983]*. Ainsi, l'huile de poisson apparaît comme diminuant la viscosité sanguine et agit favorablement sur les autres propriétés sanguines.

[15] Elle n'apparaît pas dans la bibliographie, tout comme la requête de l'étape 1. Nous avons procèdé par essais successifs jusqu'à trouver des résultats semblables à ceux décrits par Don Swanson.

Le figure 4 illustre les deux premières étapes.



D'après Swanson, 1989

Figure 4 : Méthode bibliographique, exploration

1.5.2 Seconde partie : exclusion

Etape 3 : l'hypothèse huile de poisson/maladie de Raynaud est-elle connue ? Une recherche est conduite pour vérifier qu'aucune référence bibliographique ne mentionne ensemble huile de poisson et maladie de Raynaud.

Etape 4 : déterminer si l'hypothèse de l'apport d'huile de poisson par l'alimentation pour traiter la maladie de Raynaud est médicalement recevable, en étudiant minutieusement les deux littératures.

Les étapes 3 et 4 ont été détaillées au paragraphe 1.3.2.

1.5.3 Résumé de la méthode bibliographique

Gordon et Lindsay résument ainsi la méthode employée par Swanson pour mettre à jour les liens entre l'huile de poisson et la maladie de Raynaud [Gordon, 1996] :

1 Choisir un sujet (maladie de Raynaud)

2 Rechercher la littérature **C** = {*maladie de Raynaud*}

3 Supposer que **B** (des paramètres sanguins) peut être étudié en relation avec la maladie de Raynaud

4 Rechercher la littérature **C'** = **C** $\cap$ {*sang*}

5 Relever deux concepts communs : viscosité sanguine et rigidité érythrocytaire

6 Rechercher la littérature **B'** = {*viscosité sanguine*} $\cup$ {*déformabilité érythrocytaire*}

7 Relever le concept d'huile de poisson

8 Rechercher la littérature **A** = {*huile de poisson*}

9 Montrer que {*huile de poisson*} $\cap$ {*maladie de Raynaud*} = $\varnothing$

10 Montrer et expliquer la connexion plausible entre la maladie de Raynaud et l'huile de poisson

({*X*} désigne l'ensemble des documents pour un sujet donné.)

Lors de la publication de cet article [Swanson, 1989a], cette méthodologie, qui offre la possibilité d'identifier des informations complémentaires et non liées, n'est que partiellement automatisée et comprend une partie manuelle importante. Swanson a alors pour objectif de développer un outil qui permet d'exploiter systématiquement les connexions cachées en s'appuyant sur le modèle que nous appelons ABC.

La méthodologie bibliographique est un support robuste pour formuler des hypothèses en se basant sur le savoir public caché. Elle a fait ses preuves à

travers les deux premières découvertes de Swanson. Le modèle ABC s'en inspire largement.

1.6 Le modèle ABC

Le terme "modèle ABC" est celui que nous employons par commodité. A notre connaissance, Swanson ne l'a pas utilisé.

Swanson élabora le raisonnement suivant, soit :

- A l'huile de poisson,
- B l'agrégation plaquettaire, la réactivité vasculaire et la viscosité du sang,
- C la maladie de Raynaud.

A améliore C en agissant sur B. C'est le schéma classique d'action d'un médicament. Une maladie C est caractérisée par un certain nombre de désordres physiologiques B (physiopathologie), le médicament A agit favorablement sur les désordres physiologiques.

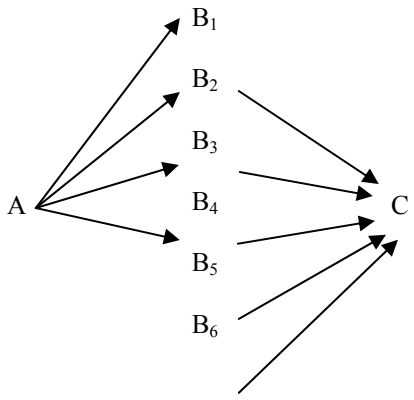


Figure 5 : Le modèle ABC

Dans ce premier exemple, Swanson connaissait les liens A→B et B→C, et $A \cap C = \varnothing$. La physiologie représente l'élément commun qui permet de lier la maladie au traitement. B peut représenter plusieurs éléments. Dans le cas de la

maladie de Raynaud, Swanson citera l'agrégation plaquettaire, la viscosité du sang et la réactivité vasculaire.

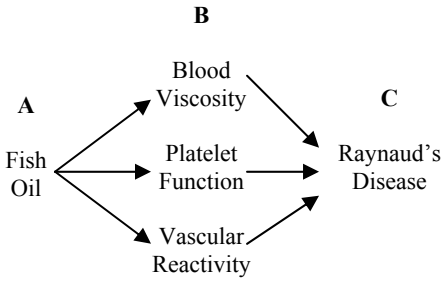


Figure 6 : Modèle ABC appliqué à l'exploration des liens entre l'huile de poisson et la maladie de Raynaud

Plus A et C ont d'éléments B communs, plus il y a de chance que le lien AC soit fort et que l'on trouve par la suite des preuves expérimentales qui valideront l'hypothèse générée.

1.6.1 Le savoir public caché

Sachant que A agit sur B (sans en être la cause exclusive) et B agit sur C, on peut formuler l'hypothèse que A agit sur C. C'est un raisonnement par inférence ou par transitivité. Si les liens AB et BC sont connus mais pas AC, Swanson parle de savoir public non découvert (*undiscovered public knowledge*) ou caché. Nous considérons habituellement que les hypothèses sont inventées et non découvertes. Cependant dans le cas où AB et BC sont connus, alors l'hypothèse "A cause C" préexiste implicitement, même si elle est inconnue, jusqu'à ce qu'on la découvre. Mettre AC en avant n'éclipse pas le fait qu'il s'agisse d'une hypothèse et que pour la valider, il faudra la confronter à l'expérimentation. L'existence de données décrivant les liens AB et AC la rendent plausible. A et C sont connectés de manière logique, mais bibliographiquement disjoint.

Les connaissances identifiées selon le modèle de Swanson ne sont qualifiées de découvertes que par défaut de lien direct AC, c'est-à-dire par l'absence

d'éléments permettant de mettre deux littératures en relation. Etant donné que l'indexation de la littérature biomédicale est incomplète, parfois imparfaite, et que la recherche de documents est aussi un processus imparfait, on ne peut justifier de la découverte d'un savoir public caché qu'en produisant une recherche documentaire dont les résultats sont négatifs.

1.6.2 Processus de découverte ouvert ou fermé

La *génération* d'hypothèses suit un processus ouvert. Le point de départ est la littérature C, connue, le but étant d'identifier B puis A, inconnus à priori. D'autres variations sont aussi possibles, comme A→B→C ou A←B→C par exemple.

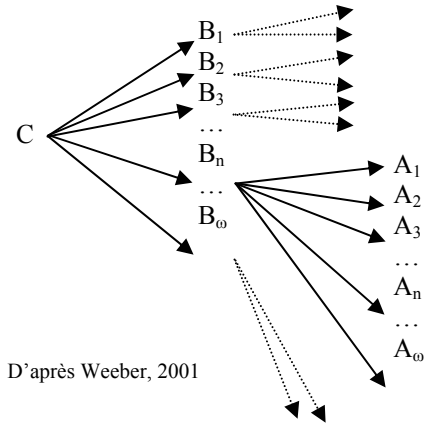


Figure 7 : Modèle ABC, processus ouvert

Les flèches pleines représentent les voies intéressantes, les pointillés des échecs.

La stratégie *trial-and-error,* plus précisément la phase exploratoire de cette stratégie, est un processus ouvert : partant de la maladie de Raynaud, Swanson recherche les dérèglements physiologiques impliqués dans la maladie, puis identifie un traitement potentiel à priori inconnu [Swanson, 1989a].

Un processus fermé *teste* une hypothèse, identifie les relations entre A et C : quels sont les liens entre la migraine et le magnésium [Swanson, 1988] ? Il est

aussi possible de tester d'autres combinaisons comme les liens entre A et B : les effets de l'arginine s'exercent-ils bien par l'intermédiaire de l'IGF I [Swanson, 1990a] ?

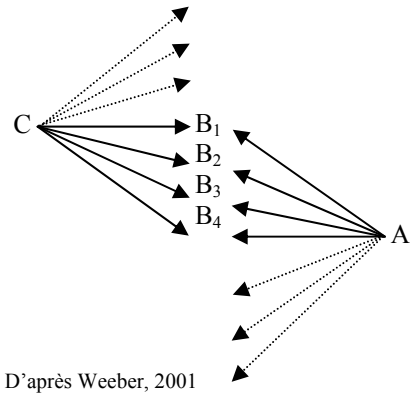


Figure 8 : Modèle ABC, processus fermé

D'un point de vue général, tester une hypothèse revient à travailler sur un volume d'information plus restreint que de générer une hypothèse. Des outils de NLP (*Natural Language Processing*) et le recours aux experts permettent de réduire le nombre de documents à traiter manuellement.

Le modèle de Swanson a servi de support à la création de plusieurs méthodologies dérivées, combinant tour à tour l'expertise et le traitement automatique de grands volumes d'information. Arrowsmith et le DAD en sont deux exemples. Ils ont été développés en ligne directe du modèle ABC, puisque la première version d'Arrowsmith est le résultat du travail de Swanson et Smalheiser, la seconde étant réalisée, entre autres par Weeber, également créateur du DAD.

1.6.3 Logique non booléenne

Le modèle ABC est non booléen, qu'une hypothèse soit générée ou testée. En se restreignant aux étapes de traitement de l'information (c'est à dire à l'exclusion

de l'expertise humaine des données) il n'y pas besoin d'avoir recours aux opérateurs booléens. Ainsi les transitions A→B et B→C servent à générer des listes de termes (B ou C) triés par fréquence ou par tout autre coefficient dans le but de mettre en avant la force des liens AB ou BC. A→B←C établit la liste des éléments B communs aux littératures A et C, éléments dont la nature peut être définie par avance, comme les phénomènes physiologiques par exemple. Là aussi, un coefficient peut permettre de pondérer la force d'un terme B en prenant en compte l'importance des liens AB et BC.

1.7 Systèmes d'aide à la découverte de connaissance

Swanson puis d'autres équipes ont tiré parti du modèle ABC pour développer des méthodes de KDD utilisant des logiciels créés sur mesure.

1.7.1 Arrowsmith

Dans les années 90, avec l'aide de Neil Smalheiser, Don Swanson a développé Arrowsmith, un système informatisé accessible librement sur Internet permettant d'explorer les liens possibles entre deux fichiers de références bibliographiques issues de Medline. Arrowsmith a pour but de guider l'utilisateur dans la découverte de relations implicites en l'aidant à sélectionner des références bibliographiques pertinentes pour établir des relations AB et BC. Ce n'est pas en lui-même un outil de découverte scientifique, mais un outil d'appui. Reprenons l'exemple de la maladie de Raynaud pour exposer rapidement le fonctionnement d'Arrowsmith.

Avant de travailler directement sur les références bibliographiques, l'utilisateur doit avoir élaboré une stratégie qui le conduit à rechercher les facteurs ou substances pouvant contribuer à traiter la maladie de Raynaud. A partir de ce point, le chercheur doit préparer ses "littératures" en utilisant Medline et décharger deux fichiers bibliographiques distincts :

1. la littérature C (fichier C), relative à la maladie de Raynaud.

2. la littérature A (fichier A), relative à une catégorie de substances ou facteurs dont on présuppose qu'ils pourraient être actifs dans la maladie de Raynaud. Dans le cas où on utilise Arroswmith pour vérifier l'hypothèse de Swanson, la littérature A est directement relative à l'huile de poisson.

En se replaçant dans les conditions de la découverte de Swanson, c'est-à-dire en 1985, il n'existe aucun lien direct entre les deux littératures. Swanson et Smalheiser conseillent d'effectuer les recherches à partir des mots des titres et de ne décharger que les titres. Lorsque ces deux fichiers sont récupérés, Arrowsmith procède en 5 étapes :

1 – Les fichiers A et C sont transmis au serveur Arrowsmith qui va créer une liste de termes et de phrases communs, la liste préliminaire des termes B. Ce processus élimine les mots vides de sens en s'appuyant sur une importante liste de mots vides.

2 – La liste préliminaire des termes B est présentée à l'utilisateur. A sa charge de la "nettoyer" en éliminant ceux qu'il ne juge pas pertinents. Il retiendra à cette étape des concepts relatifs à la viscosité sanguine et à l'agrégation plaquettaire par exemple.

3 – A partir des termes B conservés, Arrowsmith permet, pour un terme Bn donné, d'éditer la liste des titres ABn juxtaposés à ceux citant BnC afin d'aider l'utilisateur à identifier les connections AC possibles. Par exemple pour la littérature AB :

- Inhibition of platelet aggregation and thromboxane synthesis after intake of small amount of icosapentaenoic acid *[Driss, 1984]*
- The effects of cod liver oil and corn oil on platelets and vessel wall in man *[Brox, 1981]*

et pour la littérature BC :

- Assessment of platelet function in patients with Raynaud's syndrome *[Hutton, 1984]*
- Hyperviscosity and thrombotic changes in idiopathic and secondary Raynaud's syndrome *[Blunt, 1980]*

Le succès de l'opération repose sur les qualités de l'utilisateur : connaissance du sujet et de la physiologie, aptitude à imaginer ou relever les relations possibles et ingéniosité.

4 – Arrowsmith classe ensuite les termes A d'après le nombre d'associations de chaque A avec les termes B sélectionnés.

5 – Les termes A sont édités et l'utilisateur peut alors les regrouper par synonymie ou éliminer ceux qu'il ne juge pas pertinents. Puis l'étape 4 est répétée pour classer à nouveau les termes A. *Fish oil, cod liver oil* ou *eicosapentaenoic acid* apparaissent parmi les premiers termes. Si l'utilisateur a formulé l'hypothèse que l'huile de poisson peut agir sur la maladie de Raynaud par l'entremise de l'agrégation plaquettaire ou de la viscosité sanguine, il répétera alors les opérations 1 à 3, en utilisant pour littérature A des titres contenant le concept d'huile de poisson et pour littérature C des titres avec le terme Raynaud.

A aucun moment Arrowsmith vérifie si les termes A sont présents dans la littérature C. Cela peut-être fait par une simple recherche bibliographique. Ainsi, une large co-occurrence AC signifierait que le terme A retenu n'est pas porteur de nouveauté. La liste A a pour but unique de proposer à l'utilisateur une série de termes A qu'il devra examiner un par un. Après sélection d'un certain nombre de termes A, il ré-exécutera les étapes 1 à 3.

Arrowsmith peut travailler sur les titres ou les abstracts pour identifier des termes identiques dans deux fichiers différents. Sa version la plus récente s'appuie sur une liste d'environ 8000 mots-vides et est capable de traiter des termes simples (mots) ou composés (n-grammes). Relevons au passage qu'il utilise une technique de recherche non booléenne, employée dans d'autres applications, par exemple, pour des travaux de veille technologique [Dou, 1990]. Au delà de la recherche de termes identiques, Arrowsmith se repose également sur les relations de synonymies mentionnées par le MeSH, pour trouver des concepts communs. Afin de bénéficier de cette fonctionnalité, l'utilisateur doit inclure les termes du MeSH dans les citations des littératures A et C. Enfin, Arrowsmith peut retrouver certaines variations morphologiques (la plupart des variations singulier/pluriel).

Don Swanson et Neil Smalheiser ont utilisé Arrowsmith pour répéter bon nombre des travaux de Swanson [Swanson, 1997], [Smalheiser, 1998a] et [Swanson, 1999]:

- maladie de Raynaud et huile de poisson,
- migraine et magnésium,
- indométacine et maladie d'Alzheimer : effets indésirables possibles de l'indométacine par inhibition de l'action et/ou de la sécrétion d'acétylcholine,
- oestrogènes et Alzheimer : hypothèses expliquant les effets bénéfiques des oestrogènes dans la maladie d'Alzheimer,
- phospholipase A2 et schizophrénie : proposition d'un modèle animal d'étude de la régulation de la phospholipase A2.

Il existe aujourd'hui deux versions d'Arrowsmith qui, sur le fond sont identiques et sont basées sur un processus de découverte de connaissance clos, qui s'apparente plutôt à un système de validation d'hypothèse. Arrowsmith 3.0

est la version développée par Swanson au cours de années 90. Weeber et Torvik ont développé la version disponible sur le site du Département de Psychiatrie de l'Université de l'Illinois à Chicago (UIC), en collaboration avec Swanson et Smalheiser. Cette version propose une interface plus conviviale intégrant une technique de recherche différente et élimine les phases d'attentes de la version 3.0 [Smalheiser, 2002]. Quelque soit le site utilisé, il faut envoyer des données sur un ordinateur distant, ce qui pose la question de la sécurité des informations traitées et du maintient de la confidentialité autour des thèmes de recherche travaillés par les utilisateurs d'Arrowsmith.

Arrowsmith aide à mettre en lumière des liens entre deux parties connues et disjointes de la littérature biomédicale en offrant la possibilité de stimuler le raisonnement de l'utilisateur et de tester ses hypothèses. Les littératures A et C sont fixées par avance, le système ne servant qu'à proposer des liens possibles. Si la littérature A est relative à un groupe de substances (par exemple minéraux, vitamines, nutriments etc…), Arrowsmith guide l'utilisateur pour sélectionner plus précisément une ou un petit nombre de substances dans ce groupe (les huiles animales, ou l'huile de poisson).

1.7.2 Le DAD

Lindsay et Gordon [Lindsay, 1999], puis Swanson et Smalheiser [Swanson, 1997] ont proposé de réitérer l'expérience de Swanson, en utilisant l'UMLS[16] et les outils sémantiques associés. UMLS est un projet démarré en 1986 par la NLM dont le but est de faciliter le développement de systèmes informatisés "capables de comprendre" le sens de textes biomédicaux. Il regroupe les données issues de plus d'une centaine de sources (terminologies, thesaurus, classifications), tel que le MeSH ou le MeDRA (*Medical Dictionary for Regulatory Activities Terminology*) en différentes langues, ou encore *Gene*

[16] National Library of Medicine. (Page consultée le 22 septembre 2005). *Unified Medical Language System*, [En ligne]. Adresse URL : http://www.nlm.nih.gov/research/umls/umlsmain.html

Ontology[17] ou HUGO *Gene Nomenclature*[18]. UMLS est organisé en trois parties :

- le métathesaurus, vocabulaire de plus d'un million de concepts,
- le réseau sémantique, organisant les concepts selon 135 types sémantiques différents et 54 relations,
- le SPECIALIST lexicon, lexique anglais incluant les mots courants ainsi que du vocabulaire biomédical et déclinant pour chaque entrée les variations syntaxiques, morphologiques et orthographiques.

Enfin, UMLS propose toute une série de programmes informatiques pour exploiter ces trois sources, parmi lesquels MetaMap, un traducteur de texte intégral en concepts UMLS.

Weeber a mis au point le DAD, basé sur l'UMLS et le testa sur les premières hypothèses de Swanson [Weeber, 2000 et 2001]. Pour simuler la découverte de l'intérêt de l'huile de poisson dans la maladie de Raynaud, Weeber n'utilise que des références bibliographiques antérieures à novembre 1985, date des travaux de Swanson. Le DAD fonctionne en trois étapes.

1.7.2.1 Générer C→B

La littérature C, sur la maladie de Raynaud, est récupérée à partir de PubMed. La requête utilisée est "*raynaud OR raynauds*" dans le titre ou l'abstract, donnant 1.246 références. Toutes les phrases contenant le concept de "maladie de Raynaud" sont traitées avec MetaMap afin d'extraire les concepts associés. Ceux relatifs à l'anatomie et à la physiologie au sens large sont alors isolés et triés par fréquence. Le résultat de cette opération donne 57 concepts B, parmi lesquels figurent en bonne place des termes relatifs à la physiologie sanguine comme : *blood, erythrocyte deformabiblity, blood viscosity, platelet adhesiveness* et *hemorheology*. Ces termes sont sélectionnés comme concepts B.

[17] The Gene Ontology Consortium. (Page consultée le 22 septembre 2005). *Gene Ontology Home*, [En ligne]. Adresse URL : http://www.geneontology.org

1.7.2.2 Générer B→A

Partant de ces concepts B, le DAD décharge 10.611 références bibliographiques de PubMed. Les phrases où apparaissent les concepts B sont analysées pour extraire les concepts associées : 7.702 concepts sont identifiés parmi lesquels 6.747 n'apparaissant pas dans les phrases extraites de la littérature C. Ne sont alors retenu ceux relatifs aux facteurs alimentaires : vitamines, lipides, éléments, ions ou isotopes. 206 concepts sont ainsi triés par fréquence mettant en évidence l'huile de poisson (*eicosapentaenoic acid, fish oil, fatty acids omega-3, maxepa, omega-3 polyunsaturated fatty acid*). A ce point les auteurs estiment avoir les éléments nécessaires pour formuler l'hypothèse de Swanson, choisissant l'huile de poisson comme concept A.

1.7.2.3 Tester A→B←C

Le DAD permet de tester une hypothèse, en la renforçant ou en la rejetant. La littérature sur l'huile de poisson est déchargée de PubMed, ce qui représentent 463 citations. Les phrases contenant les concepts A sont soumise au même filtre sémantique que celui utilisé pour générer les concepts B à partir de la littérature C (anatomie et physiologie). Cette étape produit 45 concepts B, qui sont alors comparés aux 57 concepts B générées à l'étape 1. Les termes utilisés pour interroger PubMed à l'étape 2 sont présents, mais on trouve également des concepts supplémentaires, tels que *vasolidation, veins, capillaries, fibrinolysis, deformability, rheology* ou *dinoprostone* qui est relatif au terme générique de "réactivité vasculaire", identifiée comme importante par Swanson [Swanson, 1986a]. Ces éléments renforcent objectivement l'hypothèse d'un traitement de la maladie de Raynaud par l'huile de poisson et permettent d'orienter le travail bibliographique vers l'identification des articles complémentaires.

[18] HUGO Gene Nomenclature Committee. (Page consultée le 22 septembre 2005). *HUGO Gene Nomenclature Committee*, [En ligne]. Adresse URL : http://www.gene.ucl.ac.uk/nomenclature

1.7.2.4 Etude DAD sur de nouveaux usages potentiels de la thalidomide

La thalidomide a été commercialisée en 1957 d'abord en Allemagne, par les laboratoires Grünenthal, puis dans 46 pays, dont la plupart des pays européens. La France et les Etats-Unis ne l'ont pas autorisé. Il s'agit d'un tranquillisant mis sur le marché sans restriction spécifique alors qu'il a fait l'objet de peu d'études de toxicité. Il s'avérera tératogène et sera retiré du marché en 1962. On estime entre 8.000 à 12.000 le nombre d'enfants nés avec une malformation, dont environ 5.000 ont survécus [Silverman, 2002]. Depuis quelques années, la thalidomide est à nouveau utilisée à des fins thérapeutiques dans des conditions bien précises, en particulier, en dermatologie comme la lèpre, le syndrome de Behçet ou le lupus érythémateux discoïde [Lachapelle, 2000]. Weeber, avec l'aide d'un expert en pharmacologie et immunologie, a proposé d'autres applications possibles de la thalidomide en utilisant le DAD [Weeber, 2003]. Le lien B entre la thalidomide (A) et les maladies (C) est l'effet modulateur exercé par la thalidomide sur les cytokines. Weeber et Molema proposent l'hypothèse selon laquelle la thalidomide pourrait être un traitement de la pancréatite aiguë, de l'hépatite C chronique, de la gastrite à Hélicobacter pylori et de la myasthénie. Des recherches bibliographiques conduites sur Embase et Medline montrent qu'il existe peu ou pas de lien entre la thalidomide et ces maladies, bien que les hypothèses proposées sont basées sur un rationnel immunologique solide.

1.7.2.5 Effets indésirables désirables

Toute substance thérapeutique a des effets indésirables ou secondaires, plus ou moins graves. Dans certains cas, il arrive que ces effets non souhaités soient intéressants pour traiter une autre pathologie que celle initialement ciblée par la substance concernée. Ainsi, le sildénafil (Viagra) d'abord développé pour l'angine de poitrine et les maladies cardiaques a été réorienté vers les troubles de l'érection et a représenté 1,8 milliards de dollars de ventes pour Pfizer en 2003. Weeber propose d'utiliser le DAD pour identifier les bénéfices potentiels des

effets indésirables des médicaments. Notons au passage que *DAD signifie Drug-Adverse drug reaction-Disease* ou *Disease-Adverse drug reaction-Drug*.

Floor Rikken et Rein Vos, du *Groningen Centre for Drug Research*, ont travaillé sur les effets indésirables (ADR, *Adverse Drug Reaction*) et sur leur exploitation dans la recherche thérapeutique [Rikken, 1995a]. Ils proposent un modèle d'exploitation systématique des ADRs selon trois modalités :

1. Chimique, en s'appuyant principalement sur la relation structure-activité d'une molécule. Dans certain cas, les ADRs sont liés à la structure chimique du composé étudié. Cette information peut être mise à profit pour modifier ou dessiner de nouvelles molécules. Ainsi, la relation qui existe entre un ADR et une structure chimique donnée peut être mise à profit pour créer des molécules qui ne comporte pas cette structure.
2. Thérapeutique, l'ADR fournissant l'opportunité d'employer une molécule dans une autre pathologie que celle pour laquelle elle était initialement destinée. C'est l'exemple du sildénafil.
3. Physiopathologique, l'ADR mettant en lumière un processus physiopathologique particulier, potentiellement à l'origine de nouveaux travaux de recherche.

Rikken et Vos ont proposé une méthode bibliométrique d'exploitation des ADRs basée sur l'analyse des mots associés. Cette méthode étudie les relations entre les ADRs et les autres mots associés, en intégrant la dimension temporelle afin de détecter les évolutions de ces relations [Rikken, 1994] et [Rikken, 1995b].

1.7.3 Autres systèmes

Le modèle ABC de Swanson a inspiré un certain nombre d'autres travaux. Nous proposons de passer rapidement en revue les principaux.

A notre connaissance, Gordon et Lindsay furent les premiers à tenter de reproduire les travaux sur la maladie de Raynaud [Gordon, 1996] et la migraine [Lindsay, 1999]. Bien que leur stratégie soit semblable à celle de Swanson, cependant certains points les distinguent. En particulier, ils extraient les termes (mots simples, 2-mots ou bi-grammes et 3-mots ou tri-grammes) du texte des citations de Medline et estiment leur valeur potentielle en calculant 4 indices de statistiques lexicales pour un terme X :

- Fréquence de X au sein du corpus (occurrence)
- Fréquence de citations contenant X au sein du corpus (fréquence)
- tf × igf = occurrence de X × log (nombre de citations dans Medline/nombre de citations Medline contenant X)
- Fréquence relative = fréquence de X dans le corpus/nombre de citations Medline contenant X

Gordon et Lindsay ont développé un logiciel d'extraction de termes, éliminant les mots vides et les mots les plus fréquents dans Medline, capable de calculer les indices de statistiques lexicales. Il produit des listes de termes (n-grammes) pondérées. La sélection des termes pertinents dans les listes est réalisée "à la main". Pour les deux hypothèses, Gordon et Lindsay identifient les termes intermédiaires (B) et les substances (A) proposés par Swanson pour le traitement de la maladie de Raynaud et la migraine. Qui plus est, dans le cas de la migraine, ils identifient la cyclooxygénase à partir des données utilisée par Swanson (c'est-à-dire la littérature antérieure ou égale à 1988). L'hypothèse de l'utilisation de la cyclooxygénase, plus précisément des inhibiteurs de la COX-2, pour le traitement de la migraine n'avait alors pas encore été formulée.

Stegmann et Grohmann ont exploités l'analyse des co-occurrences pour répliquer les deux premières expériences de Swanson et proposer une hypothèse originale mentionnant l'importance du manganèse dans le déclenchement des maladies à prions [Stegmann, 2003]. Développée et exploitée par Callon l'analyse des co-occurrences se base sur le calcul de l'indice d'équivalence, qui au sein d'un corpus donné, mesure l'intensité de l'association réalisée entre deux mots :

$$Eij = \frac{C_{ij}^2}{Ci.Cj}$$

où C_{ij} est le nombre de co-occurrences des termes i et j et Ci et Cj les fréquences respectives de i et j [Callon, 1991]. Leur source d'information est Medline. Les auteurs utilisent la représentation graphique de Callon pour appréhender les littératures, en s'appuyant sur les diagrammes stratégiques et les notions de centralité (intensité des liens entre agrégats) et de densité (intensité des liens au sein d'un agrégat). Ils analysent ainsi la littérature sur la maladie de Raynaud et la littérature sur l'huile de poisson et observent que les termes intermédiaires, viscosité sanguine, agrégation plaquettaire ou vasoconstriction, se retrouvent sensiblement au même endroit sur le graphe (quadrant inférieur gauche). S'aidant de différents indices, Stegmann et Grohmann arrivent à identifier les termes intermédiaires en exploitant les termes du MeSH et les codes du champ RN (numéro de molécule chimique CAS numéro d'enzyme EC). Ils retrouvent la même distribution avec l'hypothèse migraine/magnésium.

Wren utilise la logique floue pour mesurer l'importance de la co-occurrence de deux termes de part et d'autre du lien (C→B ou B→A) à partir de Medline [Wren, 2004b]. L'extraction des concepts des littératures se fait en utilisant plusieurs sources de références. Ainsi, le MeSH est employé pour identifier les médicaments et produits chimiques, OMIM sert à identifier les phénotypes et maladie, LocusLink et *The Human Genome Nomenclature Committee* servent à identifier les gènes. Wren proposera une hypothèse originale – l'utilisation de

chlorpromazine pour réduire la progression de l'hypertrophie cardiaque – qui sera confirmé *in vivo* avec un modèle animal.

Hristovski a utilisé les règles d'association pour mesurer les liens entre deux concepts [Hristovski, 2001]. Ici encore, le MeSH, UMLS et Medline sont les sources d'information. Les règles d'associations ont été développées dans le but d'analyser le contenu des paniers des clients dans les magasins ; "si un panier contient X, il aura alors tendance à contenir également Y". Transposer ces règles aux documents revient à dire "si un document contient le concept X, il aura alors tendance à contenir le concept Y". Une citation parlant d'une maladie parlera probablement aussi de traitement. Une règle mesure la force de la co-occurrence de deux concepts dans un corpus.

D'autres équipes ont développé des techniques et des outils de KDD qui utilisent le modèle de Swanson [Gordon, 1998], [Srinivasan, 2004], [Blake, 2002], [Demaine, 2003] et [Persidis 2004]. Elles se basent toujours sur le MeSH, UMLS et Medline. Les différences résident principalement au niveau de l'extraction de l'information, dans la façon de calculer la pertinence des co-occurrences, dans la présentation des résultats et surtout par le rapport entre l'expertise humaine et l'enchaînement des opérations de traitement de l'information. Certains ont également généré de nouvelles hypothèses :

- Une défection du gène de la filamine A pourrait être à l'origine de la polymicrogyrie bilatérale périsylvienne, maladie rare où l'on observe une anomalie locale du cortex cérébral [Hristovski, 2003].
- Un lien de nature auto-immun entre les troubles bipolaires et l'arthrite rhumatoïde [Persidis, 2004].
- Le rôle bénéfique du curcuma dans la maladie de Crohn, certaines maladies de la rétine ou de la moelle épinière [Srinivasan, 2004].

1.8 Conclusion de la première partie : valeur de la méthode de Swanson

Tant qu'elles ne font pas l'objet de validations expérimentales, les hypothèses générées par la méthode de Swanson restent attaquables et prêtent le flanc à la critique. Cependant, à travers les exemples présentés dans cette première partie, deux points sont à retenir :

- Tout d'abord la méthode de Swanson a pu être utilisée pour générer d'autres hypothèses que celles relatives à la maladie de Raynaud et à la migraine. Pour mémoire :
  - lien entre les troubles bipolaires et l'arthrite rhumatoïde [Persidis, 2004],
  - bénéfice du curcuma dans la maladie de Crohn [Srinivasan, 2004],
  - nouvelles indications pour l'usage de la thalidomide [Weeber, 2003],
  - réduction de la progression de l'hypertrophie cardiaque par la chlorpromazine [Wren, 2004b],
  - une anomalie du géne de la filamine A comme cause possible de la polymicrogyrie bilatérale périsylvienne [Hristovski, 2003],
  - nous suggérons que l'adénosine diphosphate joue un rôle important dans la maladie de Raynaud, voir § 2.3.6 [Pierret, 2005].

  Il s'agit donc d'une méthode transposable à d'autres maladies, voir même d'autres thèmes. Par exemple, Swanson a classé les virus connus selon certaines caractéristiques afin de déterminer lesquels pourraient être utilisés comme armes biologiques, sur la base de sa méthode ABC [Swanson, 2001b]. La logique même du raisonnement transitif qui la soutient l'assied sur des bases solides. La transition repose sur un seul ensemble d'éléments intermédiaires. Cela n'exclue pas l'utilisation de plusieurs liens : $A \rightarrow B \rightarrow B' \rightarrow B'' \rightarrow \ldots \rightarrow C$.

- Mais plus que cette logique, le fait que plusieurs équipes différentes aient reproduit ses résultats en employant des variations bien distinctes de sa méthode, arrivant chaque fois aux mêmes conclusions, lui confère

toute sa force. Sa méthode est reproductible, indépendamment des outils employés.

Le modèle ABC, la méthodologie de Swanson et tous les outils développés autour ne servent qu'à stimuler la réflexion des scientifiques, qu'ils soient biologistes ou cliniciens. Cela représente à coup sur un avantage non négligeable en ces temps d'intense compétition. Cependant, ils ne se substituent d'aucune manière à la pensée humaine et ne sont qu'un moyen – très efficace si bien maîtrisé – d'explorer le monde des connaissances à partir de la littérature. L'ouverture d'esprit de l'expert, sa capacité à ne pas fermer certaines portes, à explorer des pistes inhabituelles, à mettre en relation des informations séparées, son aptitude à s'étonner et sa connaissance restent de précieux atouts pour réussir à tirer le meilleur parti d'une exploitation de la littérature selon Swanson. Blagosklonny et Pardee expliquent que l'exploitation par "passage en revue" (*reviewing*) des millions de données, de preuves, de faits accumulés dans les bases de données - éléments provenant de sources variées qui semblent à première vue appartenir à des domaines différents - peut générer de nouveaux savoirs [Blagosklonny, 2002]. Il s'agit de la démarche initiale de Swanson. De nouveaux savoirs peuvent être créés par l'examen de ces éléments, en les liant entre eux dans des réseaux et raisonnements vérifiables ou testables. La mise en relations d'éléments peut conduire à la formulation d'une hypothèse qui est vérifiable par l'expérience. Blagosklonny et Pardee parlent de biologie conceptuelle, qui exploite les informations des bases de données. L'analyse critique, la revue des faits et modèles existants peut produire des hypothèses dont les prédictions sont formulées de manière à pouvoir être vérifiées, le tout en utilisant des informations pertinentes provenant de publications ou rapports d'expériences dont les objectifs initiaux étaient différents – informations disjointes.

Ainsi, tout comme Swanson a formulé ses deux premières hypothèses, Ann Goodman a avancé une hypothèse originale en se basant sur l'analyse des faits

contenu dans la littérature, par recoupement d'informations [Goodman, 1998]. Elle propose que les rétinoïdes jouent un rôle important dans l'apparition de la schizophrénie en se basant sur trois faits :

1. Des anomalies congénitales semblables à celles causées par les rétinoïdes (toxicité ou déficit) sont retrouvées chez les schizophrènes.
2. Les loci[19] qui sont liées de manière théorique à la schizophrénie sont ceux des gènes de la cascade des rétinoïdes.
3. L'activation de la transcription de nombreux gènes potentiellement impliqués dans la schizophrénie sont régulés par l'acide rétinoïque.

L'hypothèse de Goodman repose sur une solide analyse de la littérature. Une rapide interrogation de Medline montre la faiblesse, voir l'inexistence, des liens entre schizophrénie et rétinoïdes avant sa première publication sur le sujet en 1994. Pour l'instant quelques d'articles de revue ont été depuis publiés, très peu relatent les résultats d'expériences. Cependant, la pertinence de l'hypothèse de Goodman est largement reconnue. Elle n'a certes pas utilisé de système de KDD, mais est arrivée au même résultat. Il reste à savoir si un tel système l'aurait, ou non, aidé à travailler plus vite ou plus efficacement. Gageons que oui.

[19] Un locus est l'emplacement physique d'un gène ou d'un groupe de gènes sur un chromosome.

Quand je définis la sérendipité comme le don de faire des trouvailles, c'est à dire de trouver ce que l'on n'a pas cherché, qu'est-ce que j'entends par trouvailles? Je parle de trouvailles si deux ou plusieurs éléments connus sont combinés originalement aux yeux de l'investigateur, en quelque chose de neuf et vrai (science), de neuf et utile (technique), ou de neuf et fascinant (art).

Pek van Andel (traduction de Danièle Bourcier)

DEUXIEME PARTIE

Le DPM
Diseases – Physiopathology - Molecules

2.1 Anamnèse

Rompant l'espace de quelques lignes avec les conventions, le "je" remplace le "nous" pour situer le contexte humain de la mise au point du DPM. Fin 2000, début 2001, je travaillais alors sur OMIM pour Fabrizio Dolfi, afin de tenter par divers tris, de mettre en évidence des informations pertinentes sur le sujet qui l'intéressait alors. Très vite, Fabrizio a compris l'importance de la bibliométrie dans l'accès à l'information. Ainsi, pendant une année, j'ai reformaté, extrait de l'information, trié, compté les mots des citations de Medline dans le but de débusquer la nouveauté. La recherche non booléenne faisait depuis longtemps parti de ma boîte à outils. Février/mars 2002, j'ai mis au point une série de "moulinettes" capables de re-traiter les descripteurs MeSH d'une notice Medline. Par exemple, l'une d'entre elles offre la possibilité de ré-indexer une notice avec des descripteurs de niveau hiérarchiques supérieurs. Dans le même temps, je scrutais Medline, à la recherche d'exemples sur lesquels m'appuyer. J'ai alors découvert l'article de Marc sur le DAD [Weeber, 2000]. *Après l'avoir lu et relu, j'ai rapidement adapté mes moulinettes MeSH et ai suivi pas à pas son mode opératoire. Fabrizio m'a apporté son expertise médicale et nous furent rapidement convaincus de la pertinence de la méthode.*

Nous avons travaillé à son amélioration. Nous baptisâmes notre outil DPM. D'un système A→B→C puis A→B←C, stimulé par les réflexions de Fabrizio, je mis au point un système non booléen fonctionnant sur plusieurs listes et capable de ne sélectionner que des concepts prédéfinis (physiologie, anatomie, protéines ...). Je pris alors contact avec Eric qui nous conseilla et me proposa de réaliser ce travail de doctorat.

2.2 Les sources de la National Library of Medicine

L'origine de la *National Library of Medicine*[20] – NLM – remonte à 1818 avec les quelques ouvrages rangés dans le bureau d'un Chirurgien Général de l'armée. Cependant, la NLM fait ses débuts en 1836 lors de la première demande officielle de budget pour l'acquisition de livre par le bureau du Général. Jusque vers 1860, cette petite collection est appelée la "*Library*". Le premier catalogue est publié en 1984. De 1865 à 1895, John Shaws Billings, assistant chirurgien se voit confier la responsabilité de développer la bibliothèque. Durant cette période, Billings fera passer la *Library* au rang de véritable bibliothèque nationale, créant *l'Index-Catalogue* et l'*Index Medicus*. La collection compte 52.000 livres ou brochures en 1876. La bibliothèque déménage en 1887 pour un nouveau bâtiment à Washington DC avec le soutien du Congrès. Elle grandit pendant plusieurs décennies et en 1956, les Sénateurs Lister Hill et John F. Kennedy soutiennent un projet de loi pour la création de la *National Library of Medicine*. C'est en 1961 que la NLM s'implante à Bethesda dans le Mariland, sur le site du *National Institute of Health*. Aujourd'hui, elle est la plus grande bibliothèque médicale au monde, recueillant des informations sur tous les thèmes relatifs à la médecine et à la biologie : soins, santé, technologie médicale, sciences sociales, physique, sciences de la vie, etc… Sa collection est riche de 7 millions d'entrées : livres, revues, rapports, manuscrits, microfilms, photographies et images, sans compter l'une des meilleures archives sur l'histoire médicale contenant d'anciens et rares travaux médicaux [Coletti, 2001] et [Knoben, 2004].

Depuis 125 ans, la NLM publie mensuellement *l'Index Medicus*, index d'articles contenus dans 4.000 revues, permettant une sélection par sujets ou auteurs. En 1960 sort la première version du *Medical Subject Headings*[21] – MeSH – vocabulaire contrôlé hiérarchisé. C'est également le début de MEDLARS, *Medical Literature Analysis and Retrieval System*. MEDLARS est l'outil

[20] National Library of Medicine. (Page consultée le 22 septembre 2005). United States *National Library of Medicine, National Institutes of Health*, [En ligne]. Adresse URL : http://www.nlm.nih.gov

[21] National Library of Medicine. (Page consultée le 22 septembre 2005). *Medical Subject Headings*, [En ligne]. Adresse URL : http://www.nlm.nih.gov/mesh/meshhome.html

informatique chargé de produire l'Index Medicus. Les capacités de recherche de MEDLARS augmentant, il devient possible de réaliser des bibliographies à la demande. MEDLARS onLINE, MEDLINE, voit le jour en 1973. C'est en juin 1997 que le Vice Président Al Gore annonce la gratuité de l'accès à Medline. PubMed[22] naît à son tour, fruit du travail du *National Center for Biotechnology Information*[23] – NCBI – dépendant de la NLM. Medline devient la principale source bibliographique de PubMed :

- 13 millions de références Medline de 1966 à aujourd'hui,
- environ 70% des références ont un résumé en anglais,
- 4.800 revues représentant 30 langues sont analysées (40 en tenant compte d'OldMedline),
- les trois quarts des références sont en anglais,
- 571.000 nouvelles références ont été ajoutées en 2004.

Medline puise la majorité de ses citations dans les revues savantes, très peu proviennent de journaux d'information généralistes ou de magazines. PubMed est intégrée au système de recherche Entrez du NCBI.

Au delà du contenu purement bibliographique de Medline, PubMed offre un accès enrichi à ses bases de données. En effet, il s'agit d'une interface qui permet de consulter Medline et OldMedline (citations avant 1966), mais également les citations en cours de catalogage/indexation ainsi que celles directement ajoutées par les éditeurs (dont certaines peuvent ensuite disparaître car de thématique trop éloigné de la médecine).

[22] NCBI/National Library of Medicine. (Page consultée le 22 septembre 2005). *PubMed*, [En ligne]. Adresse URL : http://www.pubmed.gov

[23] National Center for Biotechnology Information. (Page consultée le 22 septembre 2005). *National Center for Biotechnology Information*, [En ligne]. Adresse URL : http://www.ncbi.nlm.nih.gov

PubMed intègre aussi PubMed Central, une archive numérique ouverte et différents outils et fonctions bibliographiques comme par exemple :

- *MeSH Database*, permettant de consulter le MeSH et de construire une requête à partir de ses descripteurs.
- *Journals Database*, offrant l'accès aux revues indexées dans Medline.
- *My NCBI* (anciennement Cubby), interface personnalisée pour stocker des requêtes ou créer des DSI.
- La fonction *Related Articles*, utilisant un algorithme développé par le NCBI, qui sert à effectuer une recherche sur PubMed à partir d'un article sélectionné. *Related Articles* va identifier tous les articles semblables.
- Divers liens avec les autres bases de PubMed ou du NCBI.

Enfin, PubMed, dans son mode simple d'interrogation, traduit les requêtes des utilisateurs en utilisation UMLS, afin de leur permettre une recherche par synonymie sur le MeSH

2.2.1 La citation Medline

L'annexe 2 présente un exemple de citation Medline issue de PubMed. Très structurée et codifiée, elle permet au "bibliomètre" d'exercer son art sur de nombreux champs. Les principaux sont :

- Le titre : `TI - Medical literature as` [...]
- Le résumé d'auteur ou abstract : `AB - Specialized biomedical` [...]
- Les auteurs : `AU - Swanson DR`. Depuis peu, la NLM a ajouté le champ FAU qui écrit les prénoms des auteurs, si disponibles. Le champ AU est répété autant de fois qu'il y a d'auteurs.
- La source : `SO - Bull Med Libr Assoc 1990 Jan;78(1):29-37.`

- Les descripteurs du MeSH : `MH - Raynaud Disease/diet therapy`, que nous détaillerons dans le paragraphe suivant.
- Les CAS[24] *Registry Numbers* : `RN - 74-79-3 (Arginine)`, codes chimiques délivrés par *Chemical Abstracts Service*. Ce champ contient également les codes assignés par l'*Enzyme Commission*[25] pour désigner une enzyme ou une protéine donnée ou une famille d'enzymes ou de protéines : `RN - EC 2.7.1.37 (Protein Kinases).`

Le DPM utilise la structure d'une citation Medline pour repérer et extraire l'information du champ MH.

2.2.2 Le MeSH

Rappelons la convention d'écriture : dans les parties qui suivent, nous emploierons de nombreux descripteurs du MeSH. Convenons dès à présent de les représenter par la police de caractère `courier new`. Les requêtes PubMed seront également ainsi représentées.

2.2.2.1 Le MeSH Tree

La première liste de descripteurs a été publiée en 1954 sous le titre *Subject Heading Authority List*. Avec le début de MEDLARS en 1960, cette liste est mise à jour en profondeur et devient le MeSH, thesaurus biomédical produit par la NLM utilisé pour indexer, cataloguer et rechercher des informations et documents relatifs à la médecine, la biologie et la santé. Il est accessible en ligne à partir de plusieurs systèmes différents (la *MeSH Database*[26] de PubMed ou le

[24] American Chemical Society. (Page consultée le 22 septembre 2005). *CAS*, [En ligne]. Adresse URL : http://www.cas.org

[25] Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. (Page consultée le 22 septembre 2005). *Enzyme Nomenclature*, [En ligne]. Adresse URL : http://www.chem.qmul.ac.uk/iubmb/enzyme/

[26] NCBI/National Library of Medicine. (Page consultée le 22 septembre 2005). *MeSH*, [En ligne]. Adresse URL : http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=mesh

MeSH Browser[27]). Mis à jour chaque année, la version 2005 comporte 22.995 descripteurs, reliés entre eux par des liens de différentes natures : synonymie, équivalence, association, hiérarchie, … Le *MeSH Tree* est la forme hiérarchique du MeSH et il organise les descripteurs en 15 parties :

A - Anatomy
B - Organisms
C - Diseases
D - Chemicals and Drugs
E - Analytical, Diagnostic and Therapeutic Techniques and Equipment
F - Psychiatry and Psychology
G - Biological Sciences
H - Physical Sciences
I - Anthropology, Education, Sociology and Social Phenomena
J - Technology and Food and Beverages
K - Humanities
L - Information Science
M - Persons
N - Health Care
Z - Geographic Locations

Dans chaque catégorie, les descripteurs sont classés entre eux par thématique selon des liens hiérarchiques. Un même descripteur peut apparaître dans plusieurs branches du *MeSH Tree*. Par exemple pour Bibliometrics :

All MeSH Categories
 Information Science Category
  Information Science
   Communications Media
    Publications
     Bibliography
      Bibliometrics

All MeSH Categories
 Information Science Category
  Information Science
   Information Services
    Bibliography
     Bibliometrics

2.2.2.2 Descripteurs et subheadings

Le contexte de chaque descripteur peut être précisé par l'emploi de 83 *subheadings* ou *qualifiers*. Ainsi Bibliometrics/history signalera que l'article indexé traite de bibliométrie dans un contexte historique, ou encore

[27] National Library of Medicine. [Page consultée le 22 septembre 2005). *Medical Subject Headings*, [En ligne]. Adresse URL : http://www.nlm.nih.gov/mesh/2005/MBrowser.html

Acne Vulgaris/drug therapy mentionnera les traitements médicamenteux de l'acné alors que Acne Vulgaris/blood abordera les modifications sanguines éventuelles liées à l'acné. A titre d'exemple suivent les notices MeSH pour l'huile de poisson et la maladie de Raynaud :

Fish Oils
Oils high in unsaturated fats extracted from the bodies of fish or fish parts, especially the livers. Those from the liver are usually high in vitamin A. The oils are used as dietary supplements, in soaps and detergents, as protective coatings, and as a base for other food products such as vegetable shortenings.
Year introduced: 1983

Subheadings: *This list includes those paired at least once with this heading in MEDLINE and may not reflect current rules for allowable combinations.*
Administration and dosage, adverse effects, analysis, antagonists and inhibitors, biosynthesis, blood, chemical synthesis, chemistry, classification, diagnostic use, economics, history, immunology, isolation and purification, metabolism, pharmacokinetics, pharmacology, physiology, poisoning, radiation, effects, standards, supply and distribution, therapeutic use, toxicity, utilization.

Entry Terms:
Oils, Fish
Fish Liver Oils
Liver Oils, Fish
Oils, Fish Liver

All MeSH Categories
  Chemicals and Drugs Category
    Lipids
      Oils
        Fish Oils
          Cod Liver Oil
          Fatty Acids, Omega-3
            Docosahexaenoic Acids
            Eicosapentaenoic Acid

Raynaud Disease
An idiopathic vascular disorder characterized by bilateral Raynaud phenomenon, the abrupt onset of digital paleness or CYANOSIS in response to cold exposure or STRESS.
Year introduced: 2004 (1966)

Subheadings: *This list includes those paired at least once with this heading in MEDLINE and may not reflect current rules for allowable combinations.*
Blood, chemically induced, classification, complications, diagnosis, diet therapy, drug therapy, economics, enzymology, epidemiology, etiology, genetics, history, immunology, metabolism, microbiology, mortality, nursing, pathology, physiopathology, prevention and control, psychology, radiography, radionuclide, imaging, radiotherapy, rehabilitation, surgery, therapy, ultrasonography, urine, virology.

Entry Terms:
Raynaud's Disease
Raynauds Disease
Raynaud Phenomenon

All MeSH Categories
  Diseases Category
    Cardiovascular Diseases
      Vascular Diseases
        Raynaud Disease
          CREST Syndrome

A chaque classe de descripteurs, il est possible de combiner seulement un nombre limité de *subheadings*. En effet, si certaines combinaisons ont du sens, `Heart Diseases/surgery`, d'autres sont absurdes, `Bibliography/prevention and control`. *Entry Terms* désigne les synonymes. Certaines notices MeSH proposent la relation *See Also*, qui renvoie à des concepts voisins ou complémentaires.

Ainsi, si nous voulons faire une recherche bibliographique sur l'utilisation de l'huile de poisson contenue dans les aliments pour traiter la maladie de Raynaud, nous pourrions construire la requête :

```
"Raynaud Disease/diet therapy" AND
"Fish Oils/therapeutic use"
```

Nous retrouverions seulement 3 articles : [Swanson, 1986a, 1990] et [DiGiacomo, 1989].

2.2.2.3 Descripteurs majeurs

Lors de l'indexation, chaque descripteur est pondéré selon son importance dans l'article en question. Ainsi, dans une citation Medline, certains descripteurs peuvent être majeurs, notés avec une étoile "*". L'extrait de la référence suivante correspond à [Swanson, 1986a] :

Fish oil, Raynaud's syndrome, and undiscovered public knowledge.

```
Animals
Blood Platelets/drug effects/*physiology
Blood Vessels/drug effects/*physiology/physiopathology
Blood Viscosity/*drug effects
Fish Oils/pharmacology/*therapeutic use
Humans
Raynaud Disease/diet therapy/*physiopathology
Vascular Diseases/*prevention & control
```

A l'interrogation cela se traduit ainsi :

- Terme majeur : `"Fish Oils"[MAJR]`
- Terme majeur ou non : `"Fish Oils"[MeSH]`

2.2.2.4 Explosion : utilisation de la hiérarchie

L'utilisation du MeSH pour la recherche dans Medline offre l'option classique de la sélection des termes fils d'un concept donné. Ainsi, reprenant la hiérarchie du terme `Fish Oils`, deux choix sont possibles : soit on "explose" le terme `Fish Oils` et la recherche s'effectue non seulement sur `Fish Oils` mais aussi sur ses termes spécifiques (`Cod Liver Oil`, `Fatty Acids`, `Omega-3`, `Docosahexaenoic Acids` et `Eicosapentaenoic Acid` – les troisièmes et quatrièmes termes étant eux-mêmes fils du deuxième), soit on choisit de ne pas exploser le terme (`"Fish Oils"[MeSH :NoExp]` ou `"Fish Oils"[MAJR :NoExp]`) auquel cas, la recherche sera limitée aux citations ne contenant que ce terme. Par défaut, PubMed explose les descripteurs.

2.2.2.5 Supplementary Concepts Records

La partie la plus volumineuse du MeSH est représentée par les *Supplementary Concepts Records* (SCR), substances chimiques que les indexeurs de la NLM rencontrent dans la littérature qui alimente dans Medline. On les retrouve dans le champ RN. Au nombre de 146.248 dans la version 2005, les SCR ne sont pas des descripteurs : ils n'ont pas de contexte hiérarchique, ni de *subheadings*. Ils sont cependant rattachés à un descripteur ce qui permet une recherche assez fine, en travaillant à partir de ce descripteur désigné. Prenons l'exemple de l'adapalène, principe actif de Différine, médicament anti-acnéique. Adapalène est un SCR :

adapalene [Substance Name]

Date introduced: May 14, 1990

Registry Number: 106685-40-9

Heading Mapped to:
Naphthalenes

Entry Terms:
6-(3-(1-adamantyl)-4-methoxyphenyl)-2-naphthoic acid
CD 271
CD-271
CD271
Differin

Pharmacologic Action:
Anti-Inflammatory Agents, Non-Steroidal
Dermatologic Agents

Ce système permet de construire l'équation suivante :

```
"adapalene"[Substance Name] AND "Naphthalenes/therapeutic
 use"[MAJR:NoExp] AND "Acne Vulgaris/drug therapy"[MAJR]
```

qui sert à rechercher les citations sur l'usage d'adapalène dans le traitement de l'acné.

Dans la notice d'adapalène, la dernière rubrique *Pharmacologic Action* désigne la/les classe(s) pharmacologique(s) de la substance en question. Ce n'est pas une information exhaustive, mais consolidée. *Pharmacologic Action* regroupe à la fois des SCR et des descripteurs. La *MeSH Database* permet de construire une

équation en utilisant les classes pharmacologiques. Ainsi, `Antineoplastic Agents` contient 414 noms de substances dont 66 descripteurs et 348 SCR.

On retrouve parmi les SCR des protéines qui ne figurent pas comme descripteurs, mais qui, comme les molécules, sont rattachées à un descripteur protéine offrant la possibilité d'employer les subheadings relatif à ce descripteur.

2.2.2.6 Mises à jours du MeSH

En médecine, en biologie et dans les domaines proches, de nouveaux concepts émergent constamment, modifiant la manière dont ils s'organisent. Pour suivre ces évolutions, de nouveaux descripteurs sont ajoutés, d'autres modifiés ou effacés du MeSH, tout en tenant compte des ajustements possibles au sein de la hiérarchie. Le MeSH est mis à jour chaque année, les nouvelles versions étant disponibles vers la fin de l'automne, en même temps que les changements nécessaires sont effectués sur Medline. Le tableau 5 rapporte les évolutions entre la version 2004 et la version 2005 du MeSH.

	2004	2005
Descripteurs	22.568	22.995
SCR	~139.000	146.248
SubHeadings	83	83

Tableau 5 : Modifications du MeSH entre les versions 2004 et 2005

Pour la version 2005 :

- 487 nouveaux descripteurs sans correspondance avec des descripteurs du MeSH 2004 ont été ajoutés.
- 129 descripteurs ont été remplacés avec une terminologie plus actuelle
- 60 descripteurs ont été supprimés.

La mise à jour du MeSH est réalisée par une équipe d'une dizaine de personnes. Chacune est en charge d'un domaine précis pour lequel elle est experte. En plus des suggestions des quelques 125 personnes qui indexent les citations, l'équipe du MeSH collecte les nouveaux concepts lorsqu'ils apparaissent dans la littérature ou dans les champs de recherche émergents. Enfin, l'équipe à également recours à l'expertise externe afin de bien coordonner les différentes parties du MeSH.

2.3 La première expérience DPM

Réalisée au printemps 2002, elle s'inspire de la réplication de la première découverte de Swanson avec le DAD [Weeber, 2000]. Cette première version permet de générer une hypothèse en suivant le schéma de la figure 9.

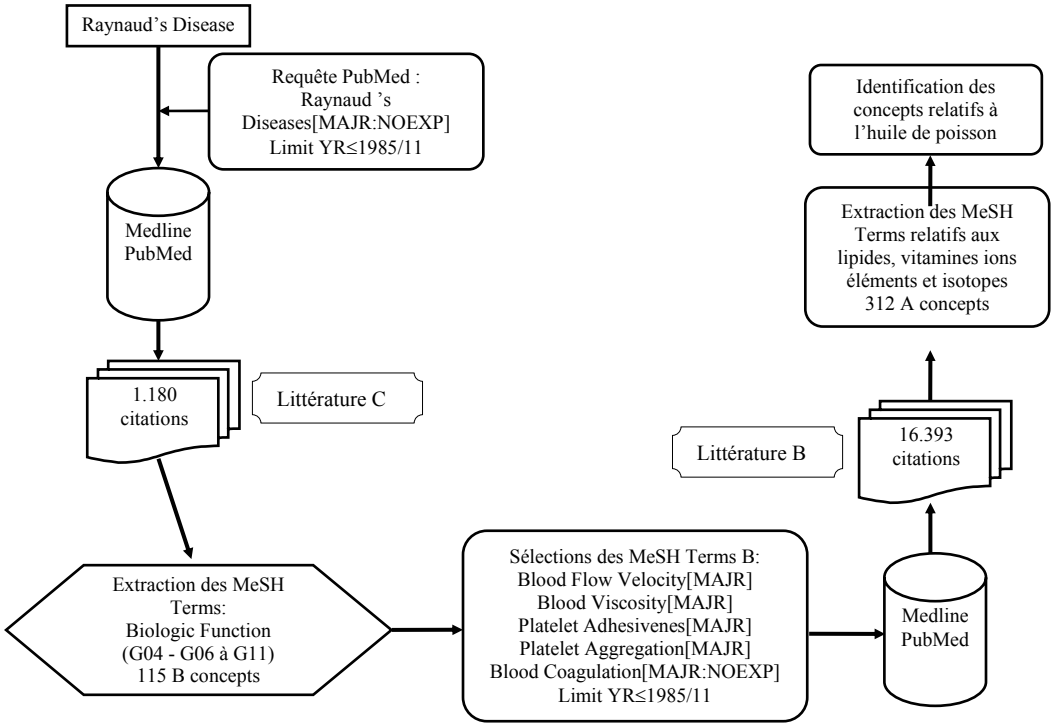


Figure 9 : Schéma de la première expérience DPM

Ces résultats sont basés sur l'exploitation de Medline le 18 décembre 2004.

Nous travaillons en suivant l'hypothèse selon laquelle lorsque deux descripteurs apparaissent ensemble dans une même référence bibliographique, il est possible qu'ils soient liés. Ce lien est d'autant plus probable que si sur un corpus, deux descripteurs apparaissent fréquemment ensemble. Par exemple, en analysant l'ensemble des descripteurs du corpus relatif à la maladie de Raynaud, il est logique que `skin temperature` ou `regional blood flow` apparaissent très souvent. Notre idée initiale fut de reproduire le DAD, en utilisant le MeSH et les descripteurs des citations. En effet, chaque article intégré dans Medline fait l'objet d'une analyse par les indexeurs de la NLM qui décrivent son contenu à l'aide d'un vocabulaire contrôlé. L'extraction de l'information sur la teneur d'un article peut donc aussi être réalisée en examinant ses descripteurs et non pas seulement par la lecture ou l'analyse NLP[28] des phrases de sa référence (titre ou abstract). Le DPM utilise le travail d'indexation réalisé tout en prenant le risque que la co-occurrence de deux descripteurs ne soit pas pertinente car la force du lien entre deux concepts est certes moindre en se plaçant au niveau de la référence bibliographique qu'au niveau de la phrase comme le fait le DAD. Nous verrons comment contourner ce problème plus loin. La base du DPM est un programme réalisé en Basic (par exemple exécutable sous l'interface Visual Basic de Microsoft Word), qui utilise une liste de référence constituée de termes du MeSH pour n'extraire d'un fichier bibliographique Medline que les descripteurs faisant partie de la liste. Le DPM génère la liste des termes retenus, trié par fréquence. Ainsi, en déchargeant de PubMed la bibliographie sur la maladie de Raynaud et en extrayant seulement les descripteurs relatifs à la physiologie, on aura une bonne idée des phénomènes que cette maladie met en jeu. Un second élément est le programme de recherche non booléenne, qui est capable d'isoler les chaînes de caractères communes à deux fichiers. Nous avons pris le parti d'expliquer le fonctionnement de notre méthode selon le processus qui a conduit à ce qu'elle est aujourd'hui, de manière chronologique. La première expérience a pour objectif de "découvrir" l'intérêt de l'huile de poisson dans le traitement de la maladie de Raynaud, en générant l'hypothèse C→B→A.

[28] Natural Language Processing

Nous avons travaillé avec un micro-ordinateur PC, processeur Pentium 4 à 2,8 GHz, avec 512 Mo de Ram, un disque dur de 29,3 Go (supportant le système d'exploitation Microsoft Windows XP, édition professionnelle 2002) et d'un disque sur réseau de 135 Go. L'accès Internet est à haut débit. Les logiciels de bases sont contenus dans la suite Microsoft Office Professionnelle. Les programmes additionnels – extraction de concepts, recherche non booléenne – ont été réalisés en Basic.

2.3.1 Constitution des dictionnaires

Le MeSH est disponible gratuitement à partir de la *MeSH Home Page*[29], après signature d'un document stipulant l'acceptation des conditions de diffusion imposées par la NLM. L'usager peut récupérer différents fichiers selon différents formats (XML ou texte). Sont proposés les notices des descripteurs, les notices des SCR, les notices des *subheadings* et le *MeSH Tree*. C'est ce dernier fichier, en texte (parce que facilement éditable) qui nous intéresse. Le *MeSH Tree* est ainsi structuré : pour chaque terme correspond un code alphanumérique qui donne sa position dans la hiérarchie. Le début du code est la lettre symbole de la partie du MeSH à laquelle appartient un descripteur donné. Pour éditer une liste de terme MeSH relatifs à la physiologie, nous procédons par interrogation du *MeSH Browser* en choisissant l'option "*Navigate from tree top*". Il est alors possible de dérouler les parties du MeSH qui sont pertinentes pour le dictionnaire que l'on souhaite construire et de soigneusement noter les codes des termes parents. Une fois cette opération réalisée, il ne reste qu'à éditer le *MeSH Tree* et à sélectionner les termes dont on a retenu les codes.

[29] National Library of Medicine. (Page consultée le 22 septembre 2005). *Medical Subject Headings*, [En ligne]. Adresse URL : http://www.nlm.nih.gov/mesh/meshhome.html

L'extrait du *MeSH Tree* qui suit provient de la partie `Information Science` et correspond au descripteur `Databases` et à ses termes fils :

```
Databases;L01.470.750
Databases, Bibliographic;L01.470.750.500
PubMed;L01.470.750.500.650
MEDLINE;L01.470.750.500.650.500
Databases, Factual;L01.470.750.750
Databases, Genetic;L01.470.750.750.325
Databases, Nucleic Acid;L01.470.750.750.325.630
Databases, Protein;L01.470.750.750.325.710
Geographic Information Systems;L01.470.750.750.462
National Practitioner Data Bank;L01.470.750.750.600
Visible Human Project;L01.470.750.750.900
```

Par habitude nous avons nommé les dictionnaires par *Tree Physiology*, *Tree Diseases*, *Tree Anatomy*, *Tree Dietary Factors*, *Tree Drugs* et *Tree Proteins/Targets*. L'annexe 3 livre en détail le contenu de chacun. Nous emploierons ces noms par la suite.

2.3.2 Interrogation de Medline sur la maladie de Raynaud

L'expérience nous a montré que pour démarrer un DPM, il faut interroger Medline en utilisant les descripteurs en tant que termes majeurs. Les bibliographies obtenues seront centrées sur le sujet qui nous intéresse et pour une citation donnée, il est très probable que tous les descripteurs soient en relation avec les descripteurs majeurs. Nous augmentons ainsi la pertinence des liens que nous mettrons en évidence. La requête utilisée est :

```
"Raynaud Disease"[MAJR:NoExp]Limits : Entrez Date to
                        1985/11
```

Nous nous sommes placés dans les conditions initiales de Swanson et avons limité par la date la sélection des références à novembre 1985 au plus tard (voir [Swanson, 1986a] page 8). Cette étape génère un fichier bibliographique ou littérature C, de 1.180 citations sur la maladie de Raynaud. Les citations sont déchargées localement d'après le format "Medline", proposé par PubMed. C'est le format de la citation présentée en annexe 2.

Depuis la création de PubMed, en 1997, le champ "EDAT- " a été ajouté aux citations Medline. Il s'agit de la date d'entrée de la notice dans PubMed. Pour les citations crées avant 1997, EDAT ou "*Entrez Date*" est égal à la date de publication de l'article. Après 1997, il s'agit de la date du jour de création de la citation dans PubMed. Ce détail est important, car non seulement la NLM indexe de nouveaux articles au fur et à mesure de leur parution – en incluant le délais de traitement bien entendu – tenant compte de ses priorités éditoriales, mais elle complète Medline avec des articles provenant de journaux publiées il y a quelques années et jusqu'alors jamais ajoutés à Medline. Il est ainsi possible qu'un article paru avant 1985 ne fût pas disponible dans Medline à cette époque, alors que 20 ans plus tard, nous y avions accès. C'est précisément le cas pour les deux citations suivantes :

- Guyotjeannin C, Lehuenen J. *Title Not Available*. Rev Hist Pharm (Paris). 1985 Mar;32(264):53-65.
- Ihde AJ. *Studies on the history of rickets. II. The roles of cod liver oil and light*. Pharm Hist. 1975;17(1):13-20.

Toutes deux furent ajoutés à Medline le 20 octobre 2001.

C'est la raison pour laquelle nous limitons nos interrogations par la date de création PubMed - "*Entrez Date*" - et non pas par la date de publication.

2.3.3 Extraction des concepts B

L'étape suivante consiste à créer la liste des descripteurs relatifs à la physiologie les plus fréquents parmi les 1.180 citations de la littérature C. Les citations de la littérature C sont nettoyées de leurs informations inutiles et reformatées : seuls les champs "MH - " sont retenus et ceux qui s'étendent sur deux lignes sont concaténés en une seule. Le programme d'extraction de terme est appliqué au fichier bibliographique en s'appuyant sur le dictionnaire *Tree Physiology* qui contient 1.505 termes MeSH relatifs à la physiologie. Cette étape isole 115 descripteurs MeSH physiologiques de la littérature C. Ce sont les termes B. La liste complète figure en annexe 4. Le tableau 6 contient les termes B de fréquence supérieure à 2.

Descripteur MeSH	Freq
Skin Temperature	92
Regional Blood Flow	86
Blood Pressure	70
Blood Viscosity	**41**
Vasoconstriction	34
Blood Flow Velocity	**32**
Biofeedback (Psychology)	30
Body Temperature	30
Blood Circulation	26
Microcirculation	26
Hemodynamic Processes	21
Vasodilation	19
Pulse	16
Body Temperature Regulation	14
Necrosis	13
Bone Resorption	11
Pain	11
Vascular Resistance	11
Heart Rate	10
Pregnancy	10
Platelet Aggregation	**8**
Fibrinolysis	7
Stress, Psychological	7
Blood Coagulation	**6**
Hematocrit	6
Reflex	6
Centromere	5
Sweating	5
Collateral Circulation	4
Differential Threshold	4
Gastrointestinal Motility	4
Laterality	4
Movement	4
Muscle Relaxation	4
Osteolysis	4
Pulmonary Diffusing Capacity	4
Systole	4
Cardiac Output	3
Chromosomes	3
Conditioning, Classical	3
Conditioning, Operant	3
Electrophysiology	3
Erythrocyte Aggregation	3
Erythrocyte Deformability	3
Evoked Potentials	3
Galvanic Skin Response	3
Muscle Contraction	3
Peristalsis	3
Platelet Adhesiveness	**3**
Posture	3
Sensation	3
Sensory Thresholds	3
Temperature Sense	3
Wound Healing	3

Tableau 6 : Première expérience DPM, concepts B, physiopathologie de la maladie de Raynaud (fréquence > 2)

Les termes utilisés par Weeber dans le DAD apparaissent tous dans cette liste, assez courte : Blood Viscosity (41), Blood Flow Velocity (32), Platelet Aggregation (8), Platelet Adhesiveness (3) et Blood Coagulation (6). Les fréquences figurent entre parenthèses. De nombreux termes complémentaires relatifs aux phénomènes hémorhéologiques impliqués dans la maladie de Raynaud et sur lesquels l'huile de poisson peut avoir une action figurent aussi dans ce tableau : Regional Blood Flow (86), Blood Pressure (70), Vasoconstriction (34), Blood Circulation (26), Microcirculation (26), Hemodynamic Processes (21), Vasodilation

(19), Pulse (16), Vascular Resistance (11), Fibrinolysis (7), Hematocrit (6), Collateral Circulation (4), Erythrocyte Aggregation (3), Erythrocyte Deformability (3), Capillary Permeability (2), Hemostasis (2), Prothrombin Time (2), Erythrocyte Count (1) et Platelet Count (1).

La somme des fréquences de tous les concepts B est de 820. La somme des fréquences des termes B relatifs à l'hémorhéologie et aux phénomènes connexes est de 430. Un peu plus de 52% des termes B sont associés aux phénomènes physiologiques identifiés par Swanson pour effectuer la transition entre la maladie de Raynaud et l'huile de poisson. Cette liste montre de manière évidente, qu'à travers la littérature biomédicale indexée dans Medline, la maladie de Raynaud est liée à des perturbations de la circulation sanguine périphérique et à des troubles de la coagulation. Notons que des termes à basses fréquences (≤ 3) sont également retenus dans cette analyse ; ces signaux faibles montrent ici toute leur importance, donnant plus de poids à des concepts de fréquences plus élevées (Prothrombin Time et Blood Coagulation, Hemostasis et Fibrinolysis, Erythrocyte Deformability et Blood Viscosity …)

2.3.4 Interrogation de Medline à partir des concepts B

Suivant le cheminement du DAD, nous avons construit la requête PubMed qui suit :

```
"Blood Flow Velocity"[MAJR] OR "Blood Viscosity"[MAJR] OR
        "Platelet Aggregation"[MAJR] OR "Platelet
  Adhesiveness"[MAJR] OR "Blood Coagulation"[MAJR:NoExp]
              Limits: Entrez Date to 1985/11
```

16.393 citations sont ainsi déchargées, il s'agit de la littérature B d'où nous allons extraire les concepts A.

Ici encore, nous avons choisi de restreindre la requête aux descripteurs majeurs.

2.3.5 Extraction des concepts A

La littérature B est reformatée de manière à ne retenir que les champs des descripteurs MeSH en les concaténant sur une seule ligne. Puis, le programme d'extraction de concepts traite les descripteurs, en utilisant le dictionnaire des termes relatifs aux vitamines, lipides, ions, éléments et isotopes (les *dietary factors* de Swanson – voir annexe A3.6). Ce dictionnaire, *Tree Dietary Factors*, comporte 541 concepts. Cette opération isole de la littérature B 312 concepts A (voir Annexe 5). Les prostaglandines figurent en tête de liste :

`Arachidonic Acids` (516)
`Epoprostenol` (467)
`Prostaglandins` (464)

et apparaissent à de très nombreuses reprises (y compris `dinoprostone`, avec une fréquence de 29). Les descripteurs relatifs à l'huile de poisson et ses principaux constituants apparaissent à partir de la 51ème position :

`Eicosapentaenoic Acid` (35)
`Fishs Oils` (14)
`Cod Liver Oils` (8)

Le poisson, en particulier l'huile de poisson, est une des sources alimentaires contenant naturellement des acides gras oméga-3. Cependant, le concept d'acide gras oméga-3 n'est pas présent dans cette liste, car ce n'est qu'en 1990 que le descripteur "`Fatty Acids, Omega-3`" a été ajouté au MeSH. Notons que "`Fatty Acids, Unsaturated`" qui est un terme générique pour les oméga-3, a une fréquence de 82.

Le tableau 7 est une sélection des concepts A relatifs aux prostaglandines et huiles naturelles.

MeSH	Freq
Arachidonic Acids	516
Epoprostenol	467
Prostaglandins	464
Lipids	321
Phospholipids	224
Prostaglandins E	220
Arachidonic Acid	214
Thromboxane B2	194
Thromboxanes	190
Thromboxane A2	166
Dietary Fats	108
Fatty Acids	92
Fatty Acids, Nonesterified	82
Fatty Acids, Unsaturated	82
Alprostadil	61
Prostaglandin Endoperoxides, Synthetic	59
Prostaglandins F	56
Prostaglandins D	51
Prostaglandins, Synthetic	49
6-Ketoprostaglandin F1 alpha	48
Prostaglandin Endoperoxides	44
Prostaglandins H	41
Linoleic Acids	36
Eicosapentaenoic Acid	35
15-Hydroxy-11 alpha,9 alpha(epoxymethano)prosta-5,13-dienoic Acid	30
Dinoprostone	29
Prostaglandins E, Synthetic	25
Oils	22
Prostaglandin D2	22
Dinoprost	20
Iloprost	17
Fatty Acids, Essential	16
Prostanoic Acids	16
Eicosanoic Acids	15
Fish Oils	14
Linolenic Acids	14
Prostaglandin H2	13
5,8,11,14-Eicosatetraynoic Acid	11
Butter	11
Prostaglandins F, Synthetic	11
Fats, Unsaturated	10
8,11,14-Eicosatrienoic Acid	9
Cod Liver Oil	8
12-Hydroxy-5,8,10,14-eicosatetraenoic Acid	7
Prostaglandins G	7
Docosahexaenoic Acids	5
Leukotrienes	5
Linoleic Acid	5
Margarine	5
Fat Emulsions, Intravenous	4
Hydroxyeicosatetraenoic Acids	4
Caproates	3
Carboprost	3
Linseed Oil	3
Castor Oil	2
Leukotriene B4	2
Oils, Volatile	2
Prostaglandins A	2
Safflower Oil	2
alpha-Linolenic Acid	1
Caprylates	1
Cottonseed Oil	1
Decanoic Acids	1
Fatty Acids, Monounsaturated	1
Fatty Alcohols	1
gamma-Linolenic Acid	1
Leukotriene A4	1
Plant Oils	1
Prostaglandins A, Synthetic	1

Tableau 7 : Première expérience DPM, concepts A, sélection de dietary factors

La somme des fréquences de tous les concepts A est de 8.150. La somme des fréquences des concepts A sélectionnés dans le tableau précédant est de 4.204, soit un peu moins de 52% des concepts A associés à la viscosité sanguine, la coagulation ou l'agrégation plaquettaire.

L'examen des concepts A confirme donc qu'un lien est possible entre l'huile de poisson et la physiopathologie de la maladie de Raynaud. Il est raisonnable de proposer l'hypothèse selon laquelle l'huile de poisson pourrait agir sur la maladie de Raynaud à travers la coagulation ou la viscosité du sang : A→B→C. A ce stade, il est possible de rechercher dans la littérature les articles complémentaires pour étayer cette hypothèse : il sera alors facile de retrouver les références sur lesquelles Swanson s'est appuyé pour son travail inaugural [Swanson, 1986a] (annexes A1.1 et A1.2).

2.3.6 Au-delà de l'huile de poisson

Avant de voir comment le DPM permet de tester l'hypothèse de Swanson, examinons ce qu'il est possible de découvrir dans les conditions de la première expérience.

En utilisant les citations ajoutées à Medline au plus tard en novembre 1985, nous avons traité les 16.393 références bibliographiques de la littérature B (voir paragraphe 2.3.4 ci-dessus), en extrayant les concepts relatifs aux médicaments et substances chimiques. Les citations sont reformatées afin de ne retenir que les descripteurs MeSH concaténés sur une seule ligne. Puis cette liste de descripteur est traitée avec notre programme d'extraction de concepts en employant le dictionnaires des substances chimique (*Tree Drugs*, voir annexe A3.1).

Nous obtenons 2.013 concepts B, liste qu'il est impossible de traiter manuellement dans sa totalité. Cependant, le terme qui apparaît avec la plus forte fréquence est *Adenosine Diphosphate*, ADP (1.752 occurrences). L'ADP est un nucléotide provenant du métabolisme de l'adénosine triphosphate (ATP). ADP et ATP sont impliqués dans différentes voies métaboliques, en particulier dans le métabolisme énergétique. La liste B fait également apparaître des

substances voisines : *Adenosine Trisphosphate* (198) et *Adenosine Nucleotides* (173). Il est établi depuis le début des années 1960 que l'ADP provoque l'agrégation plaquettaire in vitro [Hellem, 1960] et [Gaarder, 1961]. En 1975, Macfarlane propose l'hypothèse selon laquelle l'action de l'ADP sur les plaquettes serait médiée par un récepteur [Macfarlane, 1975]. La preuve est apportés par Gachet en 1992 et Fredholm précisera la nature des différents récepteurs à l'adénosine ou récepteurs purinergiques en 1994 [Gachet, 1992], [Fredholm, 1994]. Ainsi, on retrouve des récepteurs à l'adénosine sur les plaquettes (dont l'activation provoque l'agrégation) mais également sur la musculature vasculaire. A ce niveau, la stimulation des différents sous-types de récepteurs régule le tonus vasculaire, pouvant provoquer soit une vasodilatation, soit une vasoconstriction [Tabrizchi, 2001]. Il est donc évident que la combinaison des effets de l'adénosine sur les plaquettes et sur la vasculature est un facteur important de la régulation de la circulation sanguine. En particulier, l'activation du sous-type de récepteurs A2A provoque à la fois une activité anti-plaquettaire et une vasodilatation [Sandoli, 1994], [Shryock, 1997]. Un agoniste (activateur) des récepteurs A2A ou plus globalement un composé liant les récepteurs à l'adénosine pourrait alors être un traitement potentiel de la maladie de Raynaud. Ainsi, en 2002, nous avons interrogé la base de données USPTO[30] avec le terme "raynaud". De nombreux brevets revendiquent l'utilisation de composés purinergiques pour traiter la maladie de Raynaud. Le plus ancien que nous avons relevé date de 1987 (US 4,912,092). La plupart de ces brevets revendiquent l'utilisation d'antagonistes au récepteur P2YAC, pour inhiber l'agrégation plaquettaire [WO 2004052366, WO 2002098856, EP 1093814]. Toujours en 2002, l'interrogation de Pharmaprojects[31] a permis d'identifier une compagnie pharmaceutique, Aderis Pharmaceutical, qui développait alors un agoniste du récepteur A2A, le binodenoson, pour le diagnostic des maladies coronariennes, avec un positionnement possible sur la maladie de Raynaud. Nous avions formulé à posteriori cette hypothèse, avec des données antérieures

[30] United States Patent and Trademark Office. (Page consultée le 22 septembre 2005). *United States Patent and Trademark Office*, [En ligne]. Adresse URL : http://patents.uspto.gov

[31] Pharmaprojects, base de données éditée par PJB Publications rapportant le status de candidats médicaments en développement.

à 1986, date à laquelle les récepteurs purinergiques et un grand nombre des activités de l'adénosine n'étaient pas connus. Cette hypothèse doit être étayée par un travail bibliographique afin de mettre en évidence les relations potentielles qui pourront conduire à la renforcer ou à la rejeter.

2.3.7 Epilogue de la première expérience DPM

La répétition de la première expérience de Swanson en utilisant le MeSH nous a montré qu'en recherchant des connexions au niveau des références bibliographiques et non pas au niveau des phrases – comme le DAD ou Arrowsmith – on obtient également des données qui peuvent guider le chercheur vers l'hypothèse de l'utilisation de l'huile de poisson. La diversité des techniques utilisées, des champs retenus pour établir les connexions témoigne de la robustesse du principe de l'analyse. Au delà de la réplication de ce raisonnement, nous avons aussi montré que des données de 1985, comportent des connections qui prennent sens 15 à 20 ans plus tard.

Si l'on regarde de plus près les deux phases de transition, c'est-à-dire d'une part C→B, générant les concepts B et B→A, générant les concepts A, l'analyse des listes des concepts montre qu'à chaque fois près de 52% des termes des listes font en première approximation partie de thématiques voisines. Le travail sur les descripteurs MeSH s'avère donc pertinent.

Lors des toutes premières tentatives, nous avons scrupuleusement respecté le cheminement de Weeber, en éliminant des concepts A issus de la littérature B ceux qui apparaissent également dans la littérature C, afin d'être bien certain que la liste A finale est constituée de nouveautés potentielles en regard de ce qui est connu pour traiter la maladie de Raynaud. Pour la présente démonstration, nous passons cette étape sous silence. Retenons simplement, que nous avons quand même vérifié que $Fish\ Oil \cap Raynaud = \varnothing$ pour les littératures étudiées. Notre expérience nous a appris qu'en travaillant sur une maladie donnée, il est préférable de ne pas éliminer de prime abord toutes les substances décrites en relation avec elle. En effet, certaines, dont la co-occurrence avec la maladie est

basse peuvent être porteuses d'idées et donner naissance à des hypothèses prometteuses.

2.4 La deuxième expérience DPM

Après avoir généré l'hypothèse selon laquelle l'huile de poisson pourrait être un traitement de la maladie de Raynaud (C→B→A), il convient de tester cette hypothèse (C→B←A). C'est l'objectif de cette deuxième expérience dont le processus est illustré en figure 10.

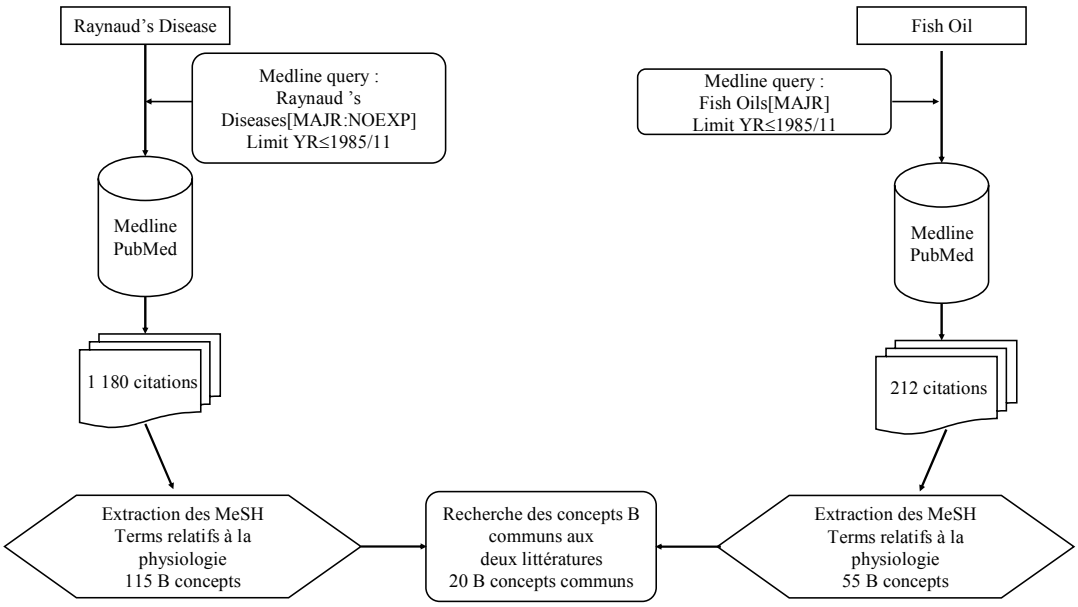


Figure 10 : Schéma de la deuxième expérience DPM

2.4.1 Extraction des concepts B de la littérature sur la maladie de Raynaud

Cette étape a été décrite au paragraphe 2.3.3 et la liste des 115 concepts B isolés lors de cette étape figure en annexe 4.

2.4.2 Extraction des concepts B de la littérature sur l'huile de poisson

La requête Medline :

```
"Fish Oils"[MAJR] Limits: Entrez Date to 1985/11
```

donne 212 citations qui sont déchargées puis reformatées pour ne retenir que les termes du MeSH. Puis cette liste de descripteurs est traitée avec notre programme d'extraction de concepts en employant le dictionnaires relatif à la physiologie (*Tree Physiology*, voir annexe A3.3). Il en ressort une liste de 55 concepts B cités dans le tableau 8. Parmi ces concepts, 13 ont un lien avec la circulation sanguine ou la coagulation (`*marqué d'une étoile`) et 3 (**`en gras`**) sont des descripteurs employés pour rechercher la littérature B dans la première expérience DPM (voir paragraphe 2.3.4).

Descripteur MeSH	Freq
***Platelet Aggregation**	24
Diet	19
Bleeding Time	9
Oxidation-Reduction	7
Pregnancy	7
***Blood Coagulation**	6
Body Weight	6
Animal Nutrition	5
*Hemostasis	4
Hydrogenation	4
Lactation	4
Species Specificity	4
*Blood Pressure	3
*Fibrinolysis	3
*Hematocrit	3
Phagocytosis	3
Structure-Activity Relationship	3
Aging	2
***Blood Viscosity**	2
*Cell Aggregation	2
Cell Survival	2
Fermentation	2
*Hemolysis	2
Membrane Fluidity	2
Necrosis	2
Oxygen Consumption	2
Alkylation	1
Bone Regeneration	1
Brain Chemistry	1
Cell Adhesion	1
Cell Membrane Permeability	1
Chemotaxis, Leukocyte	1
Drug Resistance, Microbial	1
Eating	1
Enzyme Activation	1
*Erythrocyte Count	1
*Erythrocyte Deformability	1
Gait	1
Hydroxylation	1
Infant Nutrition	1
Intestinal Absorption	1
Lymphocyte Activation	1
Metabolic Clearance Rate	1
Motor Activity	1
Muscle Contraction	1
Nutrition	1
Nutritional Requirements	1
Nutritive Value	1
Organ Size	1
Organ Specificity	1
Phenotype	1
Pinocytosis	1
*Platelet Count	1
Tissue Distribution	1
Wound Healing	1

Tableau 8 : Deuxième expérience DPM, concepts B physiologiques liés à l'huile de poisson

2.4.3 Identification des concepts B communs aux deux littératures : tester C→B←A

Notre programme de recherche non booléenne est employé pour repérer les termes B communs aux deux littératures : A sur huile de poisson et C sur maladie de Raynaud. Le résultat est traduit dans le tableau 9.

Descripteurs MeSH physiologie	Raynaud	Fish Oil
Aging	2	2
Blood Coagulation	6	6
Blood Pressure	70	3
Blood Viscosity	41	2
Body Weight	1	6
Erythrocyte Count	1	1
Erythrocyte Deformability	3	1
Fibrinolysis	7	3
Hematocrit	6	3
Hemostasis	2	4
Intestinal Absorption	1	1
Lactation	1	4
Muscle Contraction	3	1
Necrosis	13	2
Oxidation-Reduction	1	7
Oxygen Consumption	2	2
Platelet Aggregation	8	24
Platelet Count	1	1
Pregnancy	10	7
Wound Healing	3	1

Tableau 9 : Deuxième expérience DPM, concepts B communs aux littératures sur la maladie de Raynaud et sur l'huile de poisson

Il contient les 20 descripteurs MeSH relatifs à la physiologie, commun aux littératures sur l'huile de poisson d'une part et, d'autre part à la maladie de Raynaud, littératures strictement antérieures à décembre 1985. Ainsi, `Blood Pressure` apparaît 70 fois dans la littérature *Raynaud* et 3 fois dans la littérature *Fish Oil*. Ce tableau permet de contrôler les fréquences des concepts communs aux deux littératures. Il est le point de départ du raisonnement qui a pour but de tester l'hypothèse C→B←A en orientant efficacement le travail bibliographique indispensable à l'identification des articles complémentaires – mais disjoints. On retrouve trois des cinq termes qui ont servi à la sélection de la littérature B (`Platelet Aggregation`, `Blood Viscosity` et `Blood`

Coagulation), mais aussi des concepts voisins (Blood Pressure, Fibrinolysis, Hematocrit, Hemostasis, Erythrocyte Deformability, Erythrocyte Count et Platelet Count).

Pregnancy est également proposé comme étant un lien possible entre *Fish Oils* et *Raynaud Disease*. Avant 1986, la maladie de Raynaud est décrite comme étant une complication possible de la grossesse d'une part et, d'autre part, des études ont été conduites sur l'utilisation d'huile de poisson dans l'alimentation des vaches laitières. A ce niveau, sans travailler plus en avant la bibliographie, le lien huile de poisson/maladie de Raynaud à travers la grossesse n'est pas relevant. Après 1986, de nombreux articles ont été publiés sur l'impact de l'huile de poisson sur la grossesse.

2.4.4 Une première modification du tableau des concepts communs

Le tableau du paragraphe précédant ne présente pas de difficulté de lecture particulière étant court. Mais il nous est arrivé de créer des tableaux contenant plus d'une centaine d'items. Nous avons alors réfléchi au calcul d'un coefficient qui mettrait en avant les termes MeSH dont les fréquences relatives au sein des deux littératures seraient les plus hautes. La première version de ce tableau intégrant le coefficient donne ceci :

Descripteur MeSH Physiologie	Raynaud	Fish Oil	Coef.
Blood Pressure	70	3	142,45
Platelet Aggregation	8	24	130,24
Blood Viscosity	41	2	55,62
Pregnancy	10	7	47,48
Blood Coagulation	6	6	24,42
Necrosis	13	2	17,64
Fibrinolysis	7	3	14,25
Hematocrit	6	3	12,21
Hemostasis	2	4	5,43
Oxidation-Reduction	1	7	4,75
Body Weight	1	6	4,07
Aging	2	2	2,71
Lactation	1	4	2,71
Oxygen Consumption	2	2	2,71
Erythrocyte Deformability	3	1	2,04
Muscle Contraction	3	1	2,04
Wound Healing	3	1	2,04
Erythrocyte Count	1	1	0,68
Intestinal Absorption	1	1	0,68
Platelet Count	1	1	0,68
TOTAL	*182*	*81*	

Tableau 10 : Deuxième expérience DPM, introduction du coefficient

Où, pour un terme *i* donné, de fréquence *Fi.Raynaud* dans la littérature sur la maladie de Raynaud et *Fi.FishOil* dans la littérature sur l'huile de poisson :

$$Coef. = \frac{Fi.Raynaud}{\sum Fi.Raynaud / 100} \times \frac{Fi.FishOil}{\sum Fi.FishOil / 100}$$

Par exemple, pour le terme `Pregnancy` : $Coef. = \frac{10}{1,82} \times \frac{7}{0,81} = 47,48$

Le facteur 100 a été introduit pour pouvoir lire les coefficients en n'affichant que 2 décimales et ainsi contrôler que les étapes de calcul et de tri sont correctement réalisées.

Nous verrons plus loin comment nous avons généralisé le calcul du coefficient (voir § 2.5.3.1).

Le tableau réorganisé en utilisant le coefficient fait apparaître sur les trois premières lignes les termes employés pour générer la littérature B lors du passage C→B, `Blood Pressure`, `Platelet Aggregation` et `Blood Viscosity`.

Le bien-fondé et la justesse de ce coefficient peuvent certainement être discutés d'un point de vue théorique. En pratique, il sert à mettre en évidence les concepts communs à plusieurs littératures (2 ou plus), qui apparaissent le plus de fois au sein de chaque littérature prise séparément. Il ne signifie pas qu'un concept est pertinent et constitue un lien indéniable entre plusieurs éléments disjoints, mais signale que pour ce concept, la littérature est abondante. L'expert qui analyse les résultats devra ensuite vérifier s'il s'agit d'une piste pouvant conduire à un lien pertinent ou d'un concept trivial assimilable à un faux positif, donc à rejeter.

Une fois encore, le mot *expert* apparaît. Le DPM sert à générer des listes de termes potentiellement pertinents pour lier deux littératures disjointes. Il est piloté par une double expertise. D'abord une connaissance du domaine étudié, doublée d'une solide expertise en physiologie et en pharmacologie. Ensuite une expertise dans le traitement de l'information biomédicale et l'utilisation des sources telles que Medline ou le MeSH. La combinaison des deux va permettre de piloter le DPM de manière optimale et d'identifier les points sur lesquels revenir. C'est un processus itératif et dans ce cas le terme *trial-and-error* cher à Swanson n'est pas galvaudé. Il est ainsi possible d'agir sur les paramètres suivants :

- La définition des concepts pour interroger Medline
- Utilisation ou non des concepts majeurs ou de l'explosion à l'interrogation.
- Utilisation des *subheadings* à l'interrogation
- Le contenu des dictionnaires
- Interrogation sur d'autres champs : titre, abstract ou RN.

- L'extraction des termes : majeurs ou tous, avec ou sans *subheadings* particuliers

L'expérience nous a montré qu'il est préférable de modifier les paramètres un par un dans l'ordre dans lequel ils sont cités ci-dessus. Il n'y a pas de critères absolu de qualité, mais l'objectif est d'obtenir des listes relativement courtes (de quelques dizaines de termes) qui contiennent un minimum de bruit. Le bruit est estimé en tentant d'analyser un petit nombre de concepts de la liste : s'ils ne conduisent à aucune piste, ne stimulent aucune réflexion, même indirecte, nous considérons qu'en l'état il s'agit de bruit – ou que nous ne sommes pas capables d'appréhender la nature des liens que supportent ces termes. Dans un tel cas, l'essai se révèle être une impasse et, la plupart du temps, il convient de modifier les concepts employés pour interroger Medline. Le DPM s'inscrit dans une démarche itérative *trial-and-error*.

2.5 La troisième expérience DPM

Reprenant le système à deux colonnes avec coefficient, nous l'avons extrapolé à *n* colonnes où *n* représente le nombre de phénomènes physiopathologiques retenus pour une maladie donnée. Lorsque Swanson génère C→B, il ne fait ni plus ni moins que de recueillir l'ensemble des phénomènes impliqués dans C. Notre expertise, ainsi que de solides références médicales[32] nous permettent de supprimer tout ou partie de cette étape. Dans certain cas où l'on veut tester une hypothèse donnée, C→B est inutile. En revanche, générer C→B peut, dans d'autres cas stimuler la réflexion. Cette manière de procéder correspond aujourd'hui à la version la plus aboutie du DPM et se déroule en 4 étapes :

1. Choix de la pathologie à analyser, définition des phénomènes physiopathologiques à prendre en compte.
2. Traduction de la physiopathologie en requête Medline et interrogations de Medline.
3. Extraction des concepts et présentation sous forme de tableau.
4. Analyse par l'expert.

Notre troisième et dernière expérience utilise toujours l'exemple de la maladie de Raynaud.

[32] Comme par exemple :

- "Textbook of medical physiology", 2000, 10th Ed. Guyton AC & Hall JE, WB Saunders Company.
- "Goodman & Gilman's The pharmacological basis of therapeutics", 2001, 10th Ed. Hardman JG, Limbird LE & Goodman Gilman A, McGraw-Hill Medical.
- "Basic & clinical pharmacology (LANGE basic science)", 2003, 9th Ed. Katzung BG, McGraw-Hill Medical.

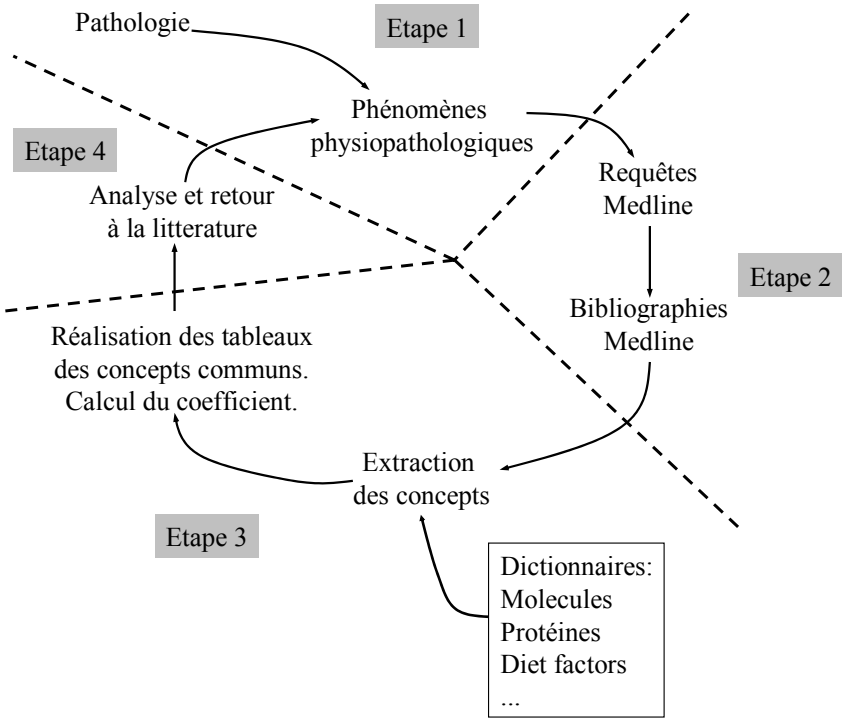


Figure 11 : Cycle du DPM

2.5.1 Etape 1 : définition de la physiopathologie

D'après le MeSH, *la physiologie est la science qui étudie le fonctionnement des êtres vivants ou de leurs organes, des facteurs physiques ou chimiques et des processus impliqués*. La physiopathologie s'intéresse aux modifications des grandes fonctions physiologiques au cours des maladies.

Ainsi, la maladie de Raynaud peut se décliner en différents aspects. Reprenant ceux qui ont retenu l'attention de Swanson, nous en avons listé quatre [Swanson, 1986a] :

- Vasoconstriction des vaisseaux irrigant les doigts.
- Agrégation plaquettaire.
- Viscosité sanguine élevée.
- Rigidité des érythrocytes.

Dans cet exemple précis, peu importe qu'un ou plusieurs de ces quatre aspects soient aujourd'hui considérés comme non pertinents. Swanson les a sélectionné compte tenu de l'état des connaissances en 1985. Nous tenons à rester dans des conditions proches de celles qui furent les siennes, prenant en considération les connaissances d'alors, ce qui pourrait s'appeler la *vérité de l'époque* – un accès au troisième monde de Popper avec les éléments dont il disposait.

Transposons ce raisonnement à une autre maladie. Quels sont les dérèglements physiologiques mis en jeu dans l'arthrite rhumatoïde ? Les auteurs travaillant sur cette pathologie s'accordent sur les phénomènes suivants [Lee, 2001] :

- Destruction synoviale.
- Activation lymphocytaire.
- Néovascularisation.
- Inflammation.
- Œdème.
- Dépôt de fibrine.

Rechercher quelles molécules peuvent agir sur chacun de ces phénomènes, indépendamment les uns des autres peut conduire à trouver de nouvelles voies pour traiter l'arthrite rhumatoïde.

Ainsi, en décrivant la pathologie en phénomènes physiopathologiques et en ne s'intéressant ensuite qu'à ces phénomènes et non plus à la maladie, nous avons "déplacé" le problème sur un champ certes plus vaste, mais qui peut aussi offrir une approche original en terme de traitement.

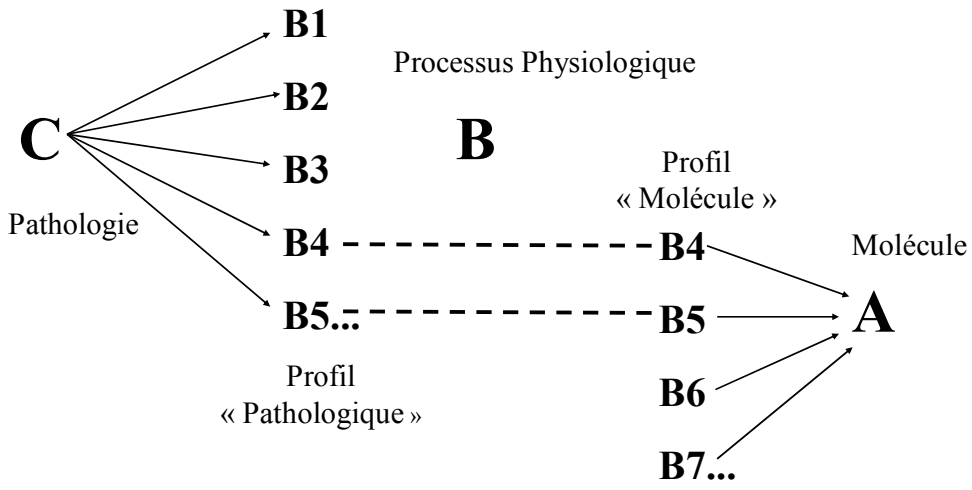


Figure 12 : Modèle ABC, centré sur l'approche DPM

Le schéma de la figure 12 peut être modifié et étendu à des liens de natures diverses entre une maladie et une molécule pour peu qu'il soit logique. Par exemple, au lieu d'observer la physiologie, il est tout à fait possible de s'intéresser aux protéines dont la régulation perturbée peut caractériser certaines maladies et sur lesquelles peuvent agir des molécules. La recherche de maladies voisines ou co-occurrentes à une maladie donnée est aussi envisageable. Ce processus peut aussi servir pour identifier, non seulement une molécule, mais une protéine ou un gène. Il peut aussi servir à ré-adresser une molécule : en décrivant l'ensemble des cibles physiologiques de cette molécule on va pouvoir rechercher quelles pathologies partagent une ou plusieurs de ces cibles. Tout est envisageable à condition que la logique du raisonnement par transitivité soit respectée, c'est-à-dire, que les liens aient du sens.

2.5.2 Etape 2 : requêtes Medline

A partir des phénomènes choisis pour définir la maladie de Raynaud, nous devons les traduire en requête pour interroger Medline. La chance est avec nous, puisque dans ce cas le MeSH nous propose un ou plusieurs termes pour chacun.

- Vasoconstriction des vaisseaux irrigant les doigts : nous recherchons des substances – *dietary factors* – qui inhibent la vasoconstriction ou provoquent une vasodilatation. Ces substances doivent agir sur le muscle lisse des vaisseaux sanguins. `Vasodilation`, `Vasoconstriction` et `Muscle, Smooth, Vascular` sont les trois termes retenus en première intention. Après différentes itérations du DPM, nous avons rejeté *Vasodilation* qui, dans ce cas, n'est pas porteur de lien.
- Agrégation plaquettaire : `Platelet Aggregation` ou `Platelet Adhesiveness` sont les deux termes MeSH à priori sélectionné. Nous retiendrons seulement `Platelet Aggregation`.
- Viscosité sanguine élevée : le descripteur correspondant est `Blood Viscosity`.
- Rigidité des érythrocytes : le descripteur correspondant est `Erythrocyte Deformability`.

Deux séries de quatre équations ont été réalisées pour illustrer l'importance du choix des descripteurs et de leur pondération en termes majeurs ou non. La première série utilise les descripteurs non pondérés. La limitation par la date est réalisée sur le champ Entrez Date. Voici les équations et leurs résultats :

`"Platelet Aggregation"[MeSH]` `Limits: Entrez Date to 1985/11`	9.949 citations

`"Muscle, Smooth, Vascular"[MESH]`
`OR "Vasoconstriction"[MESH]`
`Limits: Entrez Date to 1985/11`
	7.225 citations

`"Blood Viscosity"[MeSH]` `Limits: Entrez Date to 1985/11`	3.683 citations
`"Erythrocyte Deformability"[MeSH]` `Limits: Entrez Date to 1985/11`	251 citations

Ici, [MeSH] signifie que les concepts demandés sont recherchés en tant que termes majeurs ou non. Il n'y a pas de pondération.

Ces résultats appellent trois remarques. Tout d'abord, la quantité de données traitée : 21.108 citations, déchargées au format "Medline", qui représentent 35,9 Mo d'espace sur le disque dur. Habituellement, les opérations de bibliométrie ne traitent pas de tels volumes, qui s'apparentent tout à fait de ceux exploités en bioinformatique. Le DPM n'est pas limité par le nombre de citations à traiter. Ensuite, trois des quatre équations comportent plusieurs milliers de réponses, alors qu'une ne rapporte que 251 citations. Cette dernière pourrait être un facteur limitant dans notre analyse et nous allons comparer par la suite les résultats obtenus avec ou sans elle. Enfin, l'explosion des termes MeSH n'a pas été prise en compte puisqu'à l'exception de Muscle, Smooth, Vascular, aucuns des autres descripteurs ne possède de termes fils. Quand à Muscle, Smooth, Vascular, ce terme a un spécifique, Tunica Media, crée en 1993 et que l'on retrouve dans des citations Medline en 1992 au plus tôt. Tunica Media ne concerne donc pas l'exemple de la maladie de Raynaud.

La seconde série d'équations diffère de la première par la restriction de la recherche aux termes MeSH majeurs [MAJR].

"Platelet Aggregation"[MAJR] Limits: Entrez Date to 1985/11	4.553 citations
"Muscle, Smooth, Vascular"[MAJR] OR "Vasoconstriction"[MAJR] Limits: Entrez Date to 1985/11	3.860 citations
"Blood Viscosity"[MAJR] Limits: Entrez Date to 1985/11	1.712 citations
"Erythrocyte Deformability"[MAJR] Limits: Entrez Date to 1985/11	132 citations

Encore plus qu'avec la série d'équations précédente, Erythrocyte Deformability représente un facteur limitant par son faible nombre de citations comparé aux autres requêtes.

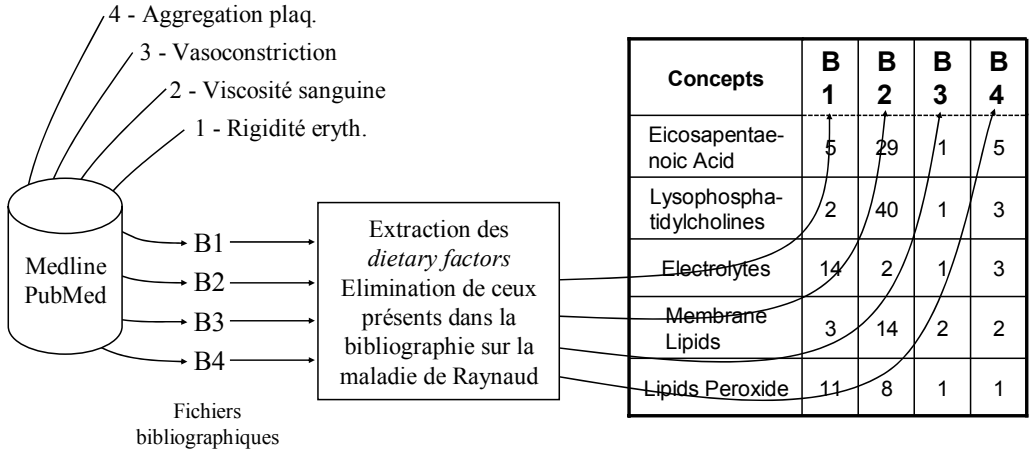
Une dernière recherche est réalisée, pour sélectionner la bibliographie sur la maladie de Raynaud, dans le but d'en extraire les *dietary factors* déjà connu dans ce contexte :

```
"Raynaud Disease"[MeSH] Limits: Entrez Date to 1985/11
```

2.129 citations sur la maladie de Raynaud sont déchargées.

2.5.3 Etape 3 : extraction des concepts, création des tableaux

Les résultats de chacune des 9 requêtes précédentes sont reformatés et traités afin de ne conserver, pour chaque fichier bibliographique, que les champs des descripteurs MeSH et de les concaténer sur une seule ligne. Les fichiers sont ensuite filtrés avec notre programme d'extraction de concepts, en employant le dictionnaire des *dietary factors* (voir annexe A3.6 *Tree Dietary Factors*). Puis, pour chacune des huit listes de *dietary factors* issues des littératures sur la vasoconstriction, l'agrégation plaquettaire, la viscosité du sang et la déformabilité des érythrocytes, on élimine les *dietary factors* déjà présents dans la littérature sur la maladie de Raynaud. Ainsi, les *dietary factors* liés à chacun des phénomènes physiologiques seront potentiellement porteurs de nouveautés car absents de la littérature sur la maladie de Raynaud.



Concepts	B 1	B 2	B 3	B 4
Eicosapentae-noic Acid	5	29	1	5
Lysophospha-tidylcholines	2	40	1	3
Electrolytes	14	2	1	3
Membrane Lipids	3	14	2	2
Lipids Peroxide	11	8	1	1

Figure 13 : Schéma de la troisième expérience DPM

Les requêtes sur les descripteurs majeurs nous permettent de construire le tableau 11, prenant en compte les quatre aspects physiopathologiques :

	Termes MeSH majeurs				
Termes MeSH *dietary factors*	**Vaso.**	**Platelet**	**Erythr.**	**Blood.**	**Coef**
Eicosapentaenoic Acid	5	29	1	5	2692,66
Lysophosphatidylcholines	2	40	1	3	1910,85
Electrolytes	14	2	1	3	1652,85
Membrane Lipids	3	14	2	2	1345,88
Lipid Peroxides	11	8	1	1	1082,78
Fish Oils	2	12	1	2	691,26

Tableau 11 : Troisième expérience DPM, résultats sur quatre littératures (requêtes avec les descripteurs majeurs)

Les abréviations employées pour tous les tableaux de cette analyse sont les suivantes :

- **Vaso.** : littérature sur la vasoconstriction ou les muscles lisses des vaisseaux sanguins.
- **Platelet.** : littérature sur l'agrégation plaquettaire.
- **Blood.** : littérature sur la viscosité du sang.
- **Erythr.** : littérature sur la déformabilité des érythrocytes.
- **Coef.** : coefficient de chaque terme MeSH calculé à partir de leurs fréquences relatives au sein de chaque littérature.

Le tableau 11 présente les *dietary factors* communs aux quatre littératures sur les phénomènes physiopathologiques associés à la maladie de Raynaud, n'apparaissant pas dans la littérature sur la maladie de Raynaud. Pour chaque terme MeSH représentant un *dietary factor* correspondent ses 4 fréquences aux seins de chacune des littératures.

L'EPA (l'acide eicosapentaenoique – constituant important de l'huile de poisson) apparaît en première position. `Fish Oils` figure aussi dans ce tableau.

2.5.3.1 Extension du calcul du coefficient à n *colonnes.*

Le principe du coefficient présenté au paragraphe 2.4.4 peut être étendu pour des tableaux à *n* colonnes. Son rôle est de pouvoir trier les termes MeSH en tenant compte de leur fréquences relatives au sein de chacune des littératures représentées dans un tableau. Le mode de calcul des fréquences relatives demeure inchangé. La fréquence relative d'un terme au sein d'une colonne est la valeur de sa fréquence rapportée au total des fréquences pour cette colonne, exprimée en pourcentage.

Nous avons adopté le raisonnement suivant pour le définir : pour un terme MeSH donné …

> … il doit prendre en compte la fréquence relative de ce terme au sein de chaque littérature.
>
> … il doit refléter l'impact global du terme (i.e. en incluant toutes les littératures).
>
> … il ne sert qu'à mettre en avant les termes pour lesquels il y proportionnellement le plus de littérature supportant un lien potentiel avec la pathologie étudiée.

Ainsi, pour un tableau à *n* colonnes, on peut imaginer une représentation géométrique plane de la valeur du coefficient d'un terme donné, par la surface qu'occuperait un polygone tracé dans un système à *n* axes ou chaque axe représente un colonne, les coordonnées sur les axes étant les fréquences relatives du terme pour chaque colonne.

Prenons l'exemple d'un terme MeSH dans tableau à 5 colonnes :

	Axe A	Axe B	Axe C	Axe D	Axe E
Terme MeSH *X*	1	2	1	3	2

Tableau 12 : Exemple de tableau DPM à 5 colonnes

Le graphe plan résultat de ce tableau pour le terme MeSH *X* est le suivant :

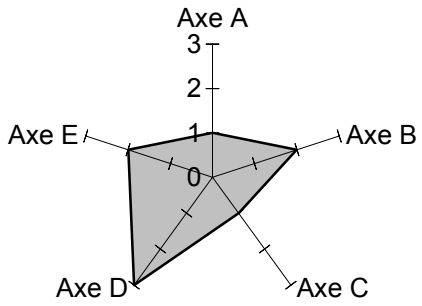


Figure 14 : Représentation graphique plane du tableau DPM à 5 colonnes

La surface grisée représente le coefficient. Pour un terme donné de fréquences relatives A, B, C, D et E :

$$Coef. = \tfrac{1}{2} \sin(360/5) \times (AB + BC + CD + DE + EA)$$

Cette équation n'est pas permutable, sauf si A=B=C=D=E. En effet, si l'ordre des colonnes est changé, la valeur du coefficient varie. Pour passer outre ce problème, nous calculons pour chaque paire possible d'axes leur contribution en terme de surface, rapporté au nombre réel d'axe :

$$Coef. = \tfrac{1}{2} \sin(360/5) \times (AB + AC + AD + AE + BC + BD + BE + CD + CE + DE) \times 5/10$$

Notre souci étant d'avoir une idée des termes soutenu par le plus de littérature au sein de l'ensemble des colonnes, dans cet exemple, nous simplifions la formule par :

$$Coef. = AB + AC + AD + AE + BC + BD + BE + CD + CE + DE$$

2.5.3.2 Autres tableaux

Trois autres tableaux ont également été générés au cours de cette expérience, en plus du tableau 11.

Tableau 13 : interrogation des quatre littératures (vasoconstriction, agrégation plaquettaire, viscosité sanguine et déformation érythrocytaire) en terme du MeSH non pondérés (i.e. majeurs ou non) et recherche des *dietary factors* communs.

	Termes MeSH non pondérés				
Termes MeSH *dietary factors*	**Vaso.**	**Platelet.**	**Erythr.**	**Blood.**	**Coef.**
Lipoproteins, LDL	78	40	1	11	1340,75
Membrane Lipids	5	37	6	11	1004,88
Lysophosphatidylcholines	3	73	2	7	537,26
Lipoproteins, HDL	40	36	1	4	470,56
Eicosapentaenoic Acid	7	60	1	7	386,87
Lipid Peroxides	25	29	3	1	381,03
Electrolytes	19	5	1	15	362,06
Cholesterol, Dietary	27	18	1	2	190,83
Rubidium	25	4	2	2	177,14
Fish Oils	2	27	1	2	82,00
Zinc	3	16	1	2	61,62
Copper	3	11	1	2	49,53
Anions	4	7	1	1	31,65

Tableau 13 : Troisième expérience DPM, résultats sur quatre littératures (requêtes avec les descripteurs non pondérés)

Tableau 14 : interrogation des trois littératures (vasoconstriction, agrégation plaquettaire et viscosité sanguine) en terme du MeSH majeurs et recherche des *dietary factors* communs.

	Termes MeSH majeurs			
Termes MeSH *dietary factors*	**Vaso.**	**Platelet.**	**Blood.**	**Coef.**
Lipoproteins, LDL	53	17	7	679,47
Lipoproteins, HDL	27	18	3	238,02
Eicosapentaenoic Acid	5	29	5	180,09
Lysophosphatidylcholines	2	40	3	125,10
Lipoproteins, VLDL	5	10	7	105,46
Electrolytes	14	2	3	53,73
Barium	22	1	2	51,01
Ascorbic Acid	13	8	1	43,58
Membrane Lipids	3	14	2	39,16
Lipid Peroxides	11	8	1	37,91
Fish Oils	2	12	2	29,67
Phosphatidylcholines	3	16	1	27,43
Lipoproteins, HDL Cholesterol	1	7	3	22,26
Cholesterol, Dietary	6	6	1	19,31
Oils	1	5	3	16,76
Niacinamide	10	2	1	16,32
Lipopolysaccharides	2	11	1	16,28
Eicosanoic Acids	1	14	1	15,96
Lipoproteins, LDL Cholesterol	3	2	1	6,07
Ions	4	1	1	5,78
Mercury	1	3	1	4,21
alpha-Tocopherol	1	2	1	3,14
Lipoproteins, VLDL Cholesterol	1	1	1	2,08

Tableau 14 : Troisième expérience DPM, résultats sur trois littératures (requêtes avec les descripteurs majeurs)

Tableau 15 : interrogation des trois littératures (vasoconstriction, agrégation plaquettaire et viscosité sanguine) en terme du MeSH non pondérés (i.e. majeurs ou non) et recherche des *dietary factors* communs.

	Termes MeSH non pondérés			
Termes MeSH *dietary factors*	**Vaso**	**Platelet**	**Blood**	**Coef**
Lipoproteins, LDL	78	40	11	185,22
Magnesium	83	54	5	139,83
Platelet Activating Factor	21	232	1	94,31
Lipoproteins, HDL	40	36	4	52,12
Electrolytes	19	5	15	42,02

Eicosapentaenoic Acid	7	60	7	41,11
Lysophosphatidylcholines	3	73	7	40,66
Membrane Lipids	5	37	11	37,31
Carbon Radioisotopes	13	60	3	29,21
Linoleic Acids	12	64	2	23,64
Lipoproteins, VLDL	9	20	8	22,61
Phosphatidylinositols	22	39	1	18,76
Cholesterol, Dietary	27	18	2	16,74
Peroxides	11	64	1	16,68
Lipid Peroxides	25	29	1	16,38
Ferricyanides	46	10	1	13,57
Barium	28	2	3	11,73
Phosphatidylcholines	5	35	3	11,71
Rubidium	25	4	2	8,32
Oils	2	19	5	8,27
Ascorbic Acid	17	14	1	6,78
Eicosanoic Acids	7	32	1	6,53
Lipoproteins, HDL Cholesterol	5	14	3	5,80
Fish Oils	2	27	2	5,00
Nitrogen	4	3	6	4,39
Lipopolysaccharides	4	17	2	4,36
Zinc	3	16	2	3,66
Niacinamide	17	4	1	3,45
Lipoproteins, LDL Cholesterol	6	6	2	2,87
Copper	3	11	2	2,75
Chylomicrons	3	2	4	2,13
Nitrites	9	2	1	1,54
Anions	4	7	1	1,41
Ions	8	2	1	1,38
Fluorides	4	5	1	1,15
Selenium	3	5	1	0,95
Cobalt	3	4	1	0,83
Vitamin A	1	2	3	0,81
Mercury	1	3	2	0,70
Carbon	2	3	1	0,55
Carbonates	1	1	2	0,40
alpha-Tocopherol	1	2	1	0,29
Osmium	1	1	1	0,21
Lipoproteins, VLDL Cholesterol	1	1	1	0,21

Tableau 15 : Troisième expérience DPM, résultats sur trois littératures (requêtes avec les descripteurs non pondérés)

Quelque soit le tableau considéré, Fish Oils et Eicosapentaenoic Acid y figurent. Apparaissent également des concepts pouvant entrer dans la composition de l'huile de poisson ou apparentés : Eicosanoic Acids, Cholesterol, Dieraty et Oils. Le nombre de concepts présents dans un

tableau diminue avec l'augmentation du nombre de colonnes, donc du nombre de littératures intermédiaires sur lesquelles le DPM est réalisé. Ce nombre diminue également si la recherche Medline sur ces littératures se fait en restreignant les requêtes aux termes MeSH majeurs. Plus on ajoute de critères, moins il y aura de réponses. Les réglages du DPM, au cours des multiples itérations réalisées avant d'atteindre un résultat sur lequel on souhaite travailler, s'effectuent d'abord sur ces deux facteurs.

Résumons les résultats obtenus dans ces 4 tableaux en examinant la position de chacun des cinq concepts que nous avons retenu :

	Vaso., Platelet., Erythr., Blood.		Vaso., Platelet., Blood.	
Tableaux …	11 Majeur	13 MeSH	14 Majeur	15 MeSH
Eicosapentaenoic Acid	1	5	3	6
Fish Oils	6	10	11	24
Cholesterol, Dietary	-	8	14	13
Eicosanoic Acids	-	-	18	22
Oils	-	-	15	20
Nb total de concepts	**6**	**13**	**23**	**44**

11 Majeur : tableau 11, interrogation en termes MeSH majeurs.
15 MeSH : tableau 15, interrogation en termes MeSH sans pondération majeur.

Tableau 16 : Récapitulatif des résultats de la troisième expérience DPM

Le tableau 11 (4 colonnes, termes majeurs) est suffisant pour orienter les recherches vers l'huile de poisson, puisque l'EPA apparaît en premier et l'huile de poisson y figure (6ème et dernière position). Nous n'avons pas examiné les possibilités offertes par les trois autres termes et sommes incapables de juger de leur pertinence comme traitement potentiel de la maladie de Raynaud, mais ce sont des concepts très proches de l'EPA ou de Fish Oils. Ainsi l'EPA est un eicosanoïde, c'est-à-dire un dérivé de l'acide eicosanoïque, acide gras en C20 – Eicosanoic Acids. Oils est le terme parent de Fish Oils. Enfin, Cholesterol, Dietary est un concept spécifique de Dietary Fats, où l'on retrouve les acides gras oméga-3 dont l'EPA.

A chacune des quatre analyses DPM, l'EPA figure toujours en haut des tableaux. L'huile de poisson se retrouve "noyée" dans la liste des concepts au fur et à mesure que les critères de recherche s'élargissent : interrogation sur 3

critères (vasoconstriction, viscosité et agrégation) et sans employer la pondération majeur.

Une remarque nous semble importante à ce niveau : l'examen des tableaux doit se faire en gardant à l'esprit que les concepts qui y figurent ne sont pas forcément tous pertinents dans l'hypothèse générée, loin s'en faut. Il nous est arrivé de ne pouvoir rien tirer de certain DPM. Prendre les concepts dans l'ordre établi par le coefficient est un moyen de gagner du temps en examinant en priorité ceux supportés par une abondante littérature. Cependant, le dernier concept d'un tableau peut s'avérer être un lien tout à fait valable. Les concepts peuvent également être regroupés par affinité. Il faut également s'attendre à ce qu'un concept par association d'idée ou par ressemblance conduise à en considérer un autre qui ne figure pas dans le tableau (stimulus à la réflexion). Enfin, il faut garder un œil à la fois critique et curieux : les concepts "qui n'ont rien à faire là" peuvent être des pistes prometteuses.

2.5.3.3 Présentation graphique

Les résultats des tableaux peuvent être présentés sous forme graphique. Ainsi, le tableau 11, en respectant l'ordre des coefficients peut ainsi être illustré :

	Termes MeSH majeurs				
Termes MeSH *dietary factors*	**Vaso.**	**Platelet**	**Erythr.**	**Blood.**	**Coef**
Eicosapentaenoic Acid	5	29	1	5	2692,66
Lysophosphatidylcholines	2	40	1	3	1910,85
Electrolytes	14	2	1	3	1652,85
Membrane Lipids	3	14	2	2	1345,88
Lipid Peroxides	11	8	1	1	1082,78
Fish Oils	2	12	1	2	691,26

Tableau 11 : Troisième expérience DPM, résultats (requêtes avec les descripteurs majeurs)

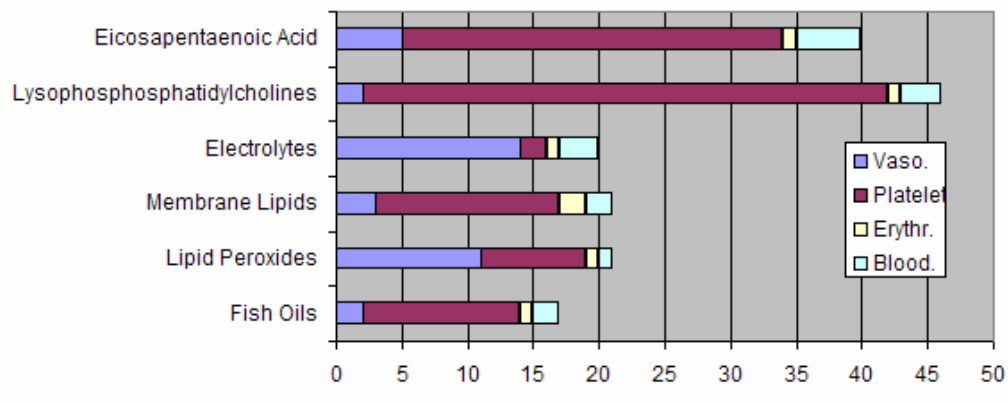


Figure 15 : Illustration graphique du tableau 11

Le graphe de la figure 15 est une aide visuelle à la lecture du tableau 11. Par exemple, la plus grande partie de la littérature est concentrée autour du thème des plaquettes, sauf pour "`Electrolytes`" et "`Lipid Peroxydes`". La présentation graphique permet de repérer plus facilement les molécules qui agissent de manière préférentielle sur un phénomène physiologique donné.

2.5.4 Etape 4 : analyse par l'expert

Un concept peut être en lui-même porteur de signification aux yeux de l'expert en charge de l'analyse. C'est une situation peu fréquente dans laquelle il a une sorte d'illumination, à l'instar du Commissaire Bourrel[33], l'évidence du lien et toutes ses implications apparaissant au grand jour.

Dans les autres cas, la plupart du temps, l'exploitation des concepts des tableaux combine l'intuition et les connaissances de l'expert avec l'exploitation de la littérature suivant quatre grandes modalités différentes selon le niveau d'évidence porté par chaque concept :

1. analyse directe de la littérature : il s'agit d'interroger Medline en croisant le concept retenu avec chacun des phénomènes physiologiques, pris séparément. Ce travail sur Medline peut – doit – être complété par

[33] "Bon sang mais c'est bien sûr !". "Euréka !" aurai-t-on dit à une autre époque.

l'utilisation d'autres sources bibliographiques (Embase, Biosis, Derwent Drug Files ou Pascal).

2. étude des ouvrages de référence : en complément de l'analyse directe de la littérature, le retour aux sources peut se révéler bénéfique et permettre de prendre de la hauteur par rapport à l'analyse. Ajoutons aux trois ouvrages de référence[34] précédemment cités, le BookShelf[35] de la NLM, qui livre sous forme électronique tout ou partie de plus d'une trentaine de livre couvrant la biologie moléculaire, la génétique, la biologie cellulaire, le cancer ou certaines autres pathologies.

3. recherche d'articles de revue de la littérature: régulièrement, les journaux scientifiques publient de tels articles – *reviews* - faisant le point sur un sujet donné. Ils sont source d'informations synthétiques souvent très utiles pour faire le point sur un sujet donné. Les articles de revue de la littérature fournissent également une bonne base pour se constituer une bibliographie sur un thème précis.

4. recherche d'informations sur l'Internet : enfin, si l'information recherchée pour établir un lien logique entre un concept et un phénomène physiologique n'est pas disponible dans la littérature ou si les éléments retrouvés sont trop ténus, l'emploi de moteurs de recherche comme Google ou Yahoo peuvent s'avérer nécessaire.

Reprenons la première ligne du tableau 11 (quatre colonnes, concepts d'interrogation en majeurs) pour illustrer brièvement la première modalité.

Termes MeSH *dietary factors*	**Vaso.**	**Platelet**	**Erythr.**	**Blood.**
Eicosapentaenoic Acid	5	29	1	5

[34] - "Textbook of medical physiology", 2000, 10th Ed. Guyton AC & Hall JE, WB Saunders Company.

- "Goodman & Gilman's The pharmacological basis of therapeutics", 2001, 10th Ed. Hardman JG, Limbird LE & Goodman Gilman A, McGraw-Hill Medical.

- "Basic & clinical pharmacology (LANGE basic science)", 2003, 9th Ed. Katzung BG, McGraw-Hill Medical.

[35] NCBI. (Page consultée le 22 septembre 2005). *Bookshelf*, [En ligne]. Adresse URL : http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books

Pour tester le lien proposé, l'EPA, l'expert doit commencer par analyser les quatre littératures suivantes :

- EPA et vasoconstriction ("Eicosapentaenoic Acid" [MeSH] AND ("Muscle, Smooth, Vascular"[MAJR] OR "Vasoconstriction"[MAJR]) Limits: Entrez Date to 1985/11) : 5 citations.
- EPA et agrégation plaquettaire ("Eicosapentaenoic Acid" [MeSH] AND "Platelet Aggregation"[MAJR] Field: All Fields, Limits: Entrez Date to 1985/11) : 29 citations.
- EPA et déformation érythrocytaire ("Eicosapentaenoic Acid" [MeSH] AND "Erythrocyte Deformability"[MAJR] Field: All Fields, Limits: Entrez Date to 1985/11) : 1 citation.
- EPA et viscosité sanguine ("Eicosapentaenoic Acid" [MeSH] AND "Blood Viscosity"[MAJR] Field: All Fields, Limits: Entrez Date to 1985/11) : 5 citations.

La lecture des titres ou des résumés de ces 39 (EPA et déformation érythrocytaire est confondue avec EPA et viscosité sanguine) citations lui montrera clairement que l'EPA a une action favorable en terme de traitement de la maladie de Raynaud sur chacun des quatre phénomènes. Relevons au passage que Swanson avait sélectionné pour sa littérature sur l'huile de poisson une partie des articles que nous retrouvons à cette étape.

EPA et …	DPM	Nb articles communs DPM et littérature sur l'huile de poisson	Références (annexe 1.2)
…vasoconstriction	5	3	*[Spector, 1983] [Morita, 1983] [Lockette, 1982]*
…agrégation	29	4	*[Von Schacky, 1985] [Driss, 1984] [Hashimoto, 1984] [Nagakawa, 1983]*
…viscosité	5	3	*[Cartwright, 1985] [Woodcock, 1984] [Terano, 1983]*
…déformabilité	1	1	*[Cartwright, 1985]*

Tableau 17 : Recoupements entre la littérature sur l'huile de poisson de Swanson (annexe 1.2) et les articles que le troisième DPM proposé pour lier l'EPA aux quatre phénomènes physiologiques

Tous ces arguments publiés sont assemblés par l'analyse DPM. Pris individuellement, ils ne sont pas porteurs de savoir qui nécessite une spécialisation extrême pour les interpréter. Ils sont du niveau de complexité des éléments que l'on retrouve dans les articles de revues et se trouvent donc à la portée d'un expert ayant une connaissance générale solide de la médecine et de la physiologie. Le DPM et, plus généralement, les méthodes de KDD basées sur le modèle de Swanson exploitent transversalement les littératures de chaque spécialité. Elles ne seraient pas viables s'il fallait recourir aux experts de chaque spécialité. Deux arguments logiquement connectés sont généralement identifiables par des experts ou des scientifiques et ce, indépendamment de leurs spécialités [Swanson, 1990b]. Relevons que dans le cadre d'une anaylse DPM la connexion de deux éléments disjoints relève d'un raisonnement par abduction, c'est-à-dire de la capacité de l'expert à générer une hypothèse plausible liant ces éléments [Demolombe, 1992]. Le raisonnement par abduction conduit à la découverte des causes et est porteur de nouvelles connaissances. L'abduction de Charles S. Peirce est proche du terme sérendipité, introduit par Horace Walpole, décrivant à l'origine la faculté de « découvrir, par hasard et sagacité, des choses qu'on ne cherche pas » [Catellin, 2003].

2.6 Les biais du DPM

Comme toute méthode, le DPM induit des biais qu'il convient de bien connaître avant toute analyse pour pouvoir mieux évaluer la valeur des résultats produits dans les tableaux. Nous en avons retenus cinq.

2.6.1 La nature du lien entre deux concepts

Le plus évident est qu'un lien est supposé exister entre deux concepts dès lors qu'ils sont co-occurrents. Or, nous l'avons vu avec les expériences de Swanson (voir 1.3.2.1), deux termes peuvent très bien figurer dans un même article sans qu'ils soient pour autant connectés de manière logique. L'analyse des concepts dans les tableaux doit tenir compte de cet aspect qui est certainement la source principale de faux positifs. Une manière de réduire ce biais pourrait être de coordonner les descripteurs entre eux à l'interrogation et à l'extraction. Par exemple, si l'on travaille sur une maladie M dans le but d'identifier les phénomènes physiologiques impliqués, il faut interroger Medline sur la maladie M en utilisant le subheading `physiopathology` (`M/physiopathology`), dans le but de ne sélectionner que des citations traitant de cet aspect. Les descripteurs des phénomènes physiologiques associés auront ainsi plus de chance d'être liés à la physiopathologie de la maladie M.

Le subheading `physiopathology` peut donc servir à cerner plus précisément les citations où un lien entre maladie et physiologie peut être décrit.

Il en va de même avec les subheadings `chemically induced` ou `drug effects` d'une part et, d'autre part, `pharmacology`, `adverse effects`, `poisoning` ou `toxicity`, qui mettent en évidence des relations entre molécules et phénomènes biologiques. L'emploi de ces couples de subheadings augmente la pertinence des références retrouvées si l'on cherche à identifier l'impact de substance sur des phénomènes biologiques [Swanson, 1990b]. Cependant, l'ajout de nouveaux critères à l'interrogation entraînera une diminution du nombre de citations déchargés : se profil le risque de "passer à côté" d'une connexion pertinente.

2.6.2 L'utilisation du MeSH

Découlant du premier biais, le second réside en l'utilisation du MeSH qui peut induire des biais de trois natures.

Le MeSH comporte certes près de 23.000 termes, mais ne décrit pas l'ensemble des concepts mis en jeux dans le domaine biomédical, domaine immense où les niveaux moléculaires, cellulaires, organes et organismes sont étroitement dépendants, mettant en jeux des relations et interactions d'une formidable complexité. Bien des aspects de la biologie ne sont pas précisément décrits par un concept du MeSH.

Se pose également la question de la nouveauté. Ainsi, pour ne prendre qu'un seul exemple, le MeSH ne décrit les différents récepteurs nucléaires de type PPARs[36] que depuis sa version 2005, alors qu'ils sont identifiés depuis plus d'une dizaine d'années. Cependant, le MeSH ne rentre pas dans la description fine des sous-types identifiés. Tout thesaurus ou vocabulaire contrôlé intègre des concepts bien établis dans la discipline à laquelle il est rattaché, laissant l'innovation et les évolutions récentes de côté en attendant qu'elles soient stabilisées, confirmées et intégrées. Il se produit un décalage entre les nouveaux concepts et ceux d'un thesaurus. Ce sont les administrateurs d'un thesaurus qui définissent le niveau de finesse que ses descripteurs vont permettre d'atteindre. Le DPM utilise le MeSH de manière détournée, utilisation qui ne figure, à notre connaissance, au cahier des charges d'aucun thesaurus.

Enfin, la constitution des dictionnaires à partir du MeSH peut elle-même introduire un biais : le DPM ne retrouvera que les concepts intégrés dans les dictionnaires employés pour une analyse donnée.

[36] *Peroxisome Proliferator-Activated Receptors*, facteurs de transcription de l'ADN. Ils sont les cibles de certains traitements du diabète de type II.

2.6.3 Choix des phénomènes physiologiques

Le contenu d'un tableau DPM est entièrement dépendant du choix des phénomènes physiologiques à traiter, en quantité et en qualité.

Plus un DPM analysera de phénomènes physiologiques simultanément, moins il mettra en évidence d'éléments communs.

Si les littératures sur les phénomènes physiologiques sont réalisées à partir de descripteurs majeurs, les liens mis en évidence seront probablement plus pertinents, mais moins nombreux. Une requête formulée avec des descripteurs majeurs aura moins de réponse que si on n'emploie pas cette pondération. Il en est de même avec l'utilisation des subheadings.

La qualité d'un tableau DPM dépend également du choix des descripteurs : trop généraux, le tableau produit sera long avec beaucoup de bruit, trop restrictifs, le tableau risque d'être vide. Réaliser différentes itérations autour d'un sujet donné, en modifiant le nombre de phénomènes physiologiques étudiés, leurs combinaisons, leurs formulations en concepts est nécessaire dans la production d'une analyse DPM.

2.6.4 Problèmes de hiérarchie

L'utilisation de concepts hiérarchisés par le DPM conduit à passer à côté de liens. Imaginons deux articles indexés l'un par un concept, l'autre par son spécifique. Ces deux articles abordent donc le même sujet (le second étant focalisé sur un point plus particulier). Ce sujet les lie. Cependant, la recherche automatique des concepts communs, pris au sens strict, n'identifiera pas ce lien. Par exemple, `Eicosapentaenoic Acid` est un spécifique de `Fish Oils`. Si ces deux descripteurs appartiennent à deux littératures différentes pour lesquelles on cherche à mettre en évidence des éléments communs, le DPM ne les retiendra pas, alors qu'ils peuvent constituer un lien pertinent. Notre outil DPM de recherche non booléenne identifie des termes, de l'enchaînement de caractères, communs à deux fichiers. Il n'interprète pas les hiérarchies du

MeSH. Retraiter les citations Medline en y ajoutant les spécifiques de chaque descripteur pourrait être un moyen de contourner ce problème.

2.6.5 Thesaurus et résultats négatifs

Les descripteurs du MeSH décrivent les concepts abordés dans un article et dans une certaine mesure, avec les subheadings, le contexte. Mais le MeSH ne permet pas de décrire la nature des résultats obtenus. Les deux propositions "A soigne B" et "A ne soigne pas B" seront décrites de la même manière. La citation dont le titre et les descripteurs suivent illustre bien ce cas :

> *Resistant arterial hypertension and hyperlipidemia: atorvastatin, not vitamin C, for blood pressure control*.
> `Antihypertensive Agents/*therapeutic use`
> `Ascorbic Acid/*therapeutic use`
> `Blood Pressure/drug effects`

`Antihypertensive Agents` correspond à l'atorvastatin. Le lien entre atorvastatin et hypertension est positif, le titre l'annonçant. En revanche, la vitamine C (`Ascorbic Acid`) ne permet pas de contrôler l'hypertension. Cette information ne figure pas dans les descripteurs. Si on ne considère que les descripteurs – ce que fait le DPM – nous pourrions créer deux co-occurrences physiologie/molécule, `Antihypertensive Agents-Blood Pressure` et `Ascorbic Acid- Blood Pressure` qui semblent indiquer que la vitamine C comme l'atorvastatine agissent sur l'hypertension. Il s'agit, pour la paire `Ascorbic Acid- Blood Pressure` d'un faux positif, que l'expert détectera lors de l'analyse.

2.7 Conclusion de la deuxième partie

La force du DPM ne réside pas tant dans la technique employée pour traiter la bibliographie mais bien dans sa modularité. Avant toute expérience, il faut avoir défini la question à laquelle le DPM doit tenter d'apporter une réponse - ou au moins quelques éléments de réponse. Puis construire l'outil pour y répondre : choix des mots-clés et des dictionnaires, modes d'interrogations, choix des colonnes, générer ou tester une hypothèse. Bien conscient des biais inhérents à son fonctionnement, l'expert pourra alors jouer sur ces différents paramètres pour affiner les résultats obtenus. Parfois en vain. Nous n'avons, par exemple, pas encore réussi à identifier de liens entre le minoxidil[37] et la pousse des cheveux. D'autre fois avec succès, comme le montre l'exemple de la maladie de Raynaud, que nous avons répliqué en identifiant non seulement l'huile de poisson, mais également l'ADP.

En écho de la première partie, notre travail à partir de la méthodologie de Swanson aboutit à la même conclusion : l'huile de poisson est un traitement potentiel de la maladie de Raynaud. Il s'agit là d'une preuve de plus en faveur du modèle ABC : sur un thème donné, peu importe la technique utilisée pour l'appliquer, les hypothèses générées seront semblables (voir § 1.7.3).

[37] Le minoxidil est un antihypertenseur utilisé pour le traitement de l'hypertension sévère. Il est également employé pour traiter l'alopécie androgénétique.

The key discovery was Jim's determination of the exact nature of the two base pairs (A with T, G with C). He did this not by logic, but by serendipity… In a sense Jim's discovery was luck, but then most discoveries have an element of luck in them. The more important point is that Jim was looking for something significant and immediately recognized the significance of the correct pairs when he hit upon them by chance.

Francis Crick

TROISIEME PARTIE

Evolutions possibles du DPM

Comme nous l'avons vu au paragraphe 1.7.3, il existe plusieurs autres systèmes de découverte de connaissances, basées sur différentes méthodes, exploitant diverses sources et outils : bases de données bibliographiques, thesaurus, bases factuelles, nomenclatures ou ontologies. En utilisant Medline comme source de données, le DPM peut s'affranchir du MeSH pour travailler sur d'autres champs que celui des descripteurs : les mots du titre ou du résumé ou les codes du champ RN. D'autres sources peuvent également être utilisées telles que des bases bibliographiques comme Biosis ou Embase, ou des bases factuelles, OMIM par exemple. Internet est aussi une source d'information exploitée en KDD.

3.1 Le DPM et Medline

Deux types de champs peuvent être mis à profit : les champs de texte libre, non contrôlés et les champs de codes. Les exemples qui suivent ressemblent techniquement au DPM, au moins dans la manière de traiter les résultats de la recherche non booléenne. Ainsi, les résultats obtenus sont présentés sous forme

de tableau. Par contre, nous n'avons pas exploité le MeSH pour sélectionner les mots ou termes des différentes listes générées selon leur type sémantique, comme nous le faisons "classiquement" avec le DPM en ne nous intéressant qu'aux termes désignant la physiologie par exemple.

3.1.1 DPM et texte libre

Le titre et le résumé sont les deux champs sur lesquelles s'appuient la majorité des méthodes de découverte de connaissances lorsqu'elles utilisent des bases de données bibliographiques. Swanson a mené à bien ses recherches bibliographiques sur l'huile de poisson et la maladie de Raynaud en travaillant sur les titres, les résumés et les descripteurs [Swanson, 1986a]. Par exemple, générer une hypothèse nouvelle sur l'utilisation d'une molécule dans le traitement d'une maladie donnée peut être réalisée en examinant les titres des littératures. C'est un avantage de la littérature biomédicale : les titres des articles sont très souvent spécifiques, précis, porteurs d'information et suggèrent des relations de causes à effets. Un titre signale les connections logiques abordées dans l'article [Swanson, 1989b].

Voici les résultats de notre travail sur les titres, d'une part, et sur les titres et les abstracts d'autre part, en traitant les corpus documentaires employés pour les première et deuxième expériences DPM (§ 2.3.2 et 2.4). Nous rappelons ici les équations employées pour générer les fichiers bibliographiques :

```
"Fish Oils"[MAJR] Limits: Entrez Date to 1985/11
```

(212 citations)

```
"Raynaud Disease"[MAJR:NoExp] Limits : Entrez Date to
1985/11
```

(1 180 citations)

3.1.1.1 Travail sur les titres seuls

Les deux fichiers bibliographiques sont traités de la même manière :

1. La casse des lettres est basculée en majuscules afin d'éviter les problèmes lors des comptages et tris.
2. Tous les signes et caractères de ponctuation sont éliminés (. , ; : / ( ) - ? [ ] = + « » ' etc …).
3. La fréquence ce chaque mot est calculée.

227 mots communs aux deux fichiers sont identifiés et présentés dans un tableau identique à celui employé pour le DPM (deuxième expérience). Puis, les mots vides de sens ou sans information sont éliminés et les pluriels sont convertis en singuliers, réduisant le nombre de mots pertinents à 59. Le coefficient est également calculé, selon la formule présentée au § 2.4.4, en utilisant les occurrences des mots retenus. Nous obtenons le tableau 18.

Mots des titres	Raynaud	Fish Oil	Coef.
BLOOD	64	10	0,63
PLATELET	11	25	0,41
PLASMA	16	12	0,35
LIVER	2	64	0,28
OIL	1	119	0,27
PROSTAGLANDIN	14	8	0,26
ARTERIAL	31	3	0,24
VASCULAR	70	1	0,21
PROSTACYCLIN	11	4	0,17
THROMBOXANE	6	7	0,16
VISCOSITY	18	2	0,15
SCLEROSIS	34	1	0,15
LUPUS	14	2	0,13
SKIN	25	1	0,12
DIET	2	12	0,12
PRESSURE	10	2	0,11
LIPIDS	1	19	0,11
AGGREGATION	3	5	0,10
ARTHRITIS	5	3	0,10
ERYTHROCYTE	3	5	0,10
CAPILLARY	15	1	0,10
HEART	3	4	0,09
SERUM	1	11	0,08
CALCIUM	5	2	0,08
ATHERO-SCLEROSIS	4	2	0,07
BLEEDING	2	3	0,06
IRON	2	3	0,06
NECROSIS	5	1	0,06
LIPOPROTEINS	1	5	0,06
AUTOIMMUNE	4	1	0,05
RHEUMATOID	4	1	0,05
CEREBRAL	3	1	0,04
CHANNEL	3	1	0,04
CLEARANCE	3	1	0,04
COLLAGEN	3	1	0,04
DEFORMABILITY	3	1	0,04
DYSTROPHY	3	1	0,04
INFARCTION	3	1	0,04
MUSCULAR	3	1	0,04
ANGINA	2	1	0,04
HEALING	2	1	0,04
HYPERPLASIA	2	1	0,04
IMMUNO-LOGICALLY	2	1	0,04
LUNG	2	1	0,04
VESSEL	1	2	0,04
AORTIC	1	1	0,02
COAGULATION	1	1	0,02
DIABETIC	1	1	0,02
DL-ALPHA-TOCOPHERYL	1	1	0,02
FIBRINOLYSIS	1	1	0,02
HYPERTENSIVE	1	1	0,02
ISCHAEMIC	1	1	0,02
LEUKOCYTES	1	1	0,02
MEDITERRANEAN	1	1	0,02
NORWEGIAN	1	1	0,02
OXYGEN	1	1	0,02
PEPTIDES	1	1	0,02
TEMPERATURES	1	1	0,02
TUMORS	1	1	0,02

Tableau 18 : DPM sur les mots des titres

Toute une série de mots relatifs à l'hémorhéologie se retrouvent dans ce tableau : *blood*, *platelet*, *arterial*, *vascular*, *viscosity*, *aggregation*, *erythrocyte*, *capillary*, *deformability*, *vessel*, *coagulation* et *fibrinolysis*. Remarquons également la présence de *prostaglandin*, *prostacyclin* et *thromboxane*, trois groupes de composés de la famille des eicosanoïdes (acides gras insaturés), dont l'EPA fait également partie :

- les prostaglandines sont impliquées dans de multiples voies de signalisation cellulaires. En particulier, la prostacycline (ou prostaglandine I2), est connue comme étant un puissant vasodilatateur et

un inhibiteur de l'agrégation plaquettaire *[Moncada, 1983]*, *[Moncada, 1984]*, *[Weksler, 1984]*.

- les molécules du groupe des thromboxanes provoquent la contraction des artères et l'agrégation des plaquettes *[Hung, 1982]*.

L'ensemble de ces mots suggère fortement qu'il existe des liens entre l'huile de poisson et la maladie de Raynaud : les plaquettes, la viscosité sanguine et le tonus vasculaire.

3.1.1.2 Travail sur les titres et abstracts

Le traitement des fichiers est le même que celui appliqué aux titres, en incluant cette fois les mots des abstracts. 1382 mots communs aux deux fichiers sont identifiés. Nous avons éliminé les mots vides de sens et sans intérêt dans le contexte de notre étude et transformé les pluriels en singulier. Restent au final 148 mots des titres et abstracts, avec les occurrences desquels nous calculons le coefficient afin d'établir le tableau 19.

Mots des titres ou abstracts	Ray.	Fish Oil	Coef.
BLOOD	299	31	14,74
PLATELET	56	127	11,31
PLASMA	80	70	8,91
ARTERY	271	14	6,04
TEMPERATURE	175	8	2,23
PRESSURE	115	11	2,01
ARTERIAL	114	11	1,99
PROSTAGLANDIN	31	39	1,92
ACID	5	223	1,77
VASCULAR	168	6	1,60
AGGREGATION	19	35	1,06
OIL	2	330	1,05
VISCOSITY	71	9	1,02
THROMBOXANE	27	22	0,94
SKIN	114	5	0,91
ISCHEMIA	55	10	0,87
SERUM	11	39	0,68
COLLAGEN	21	19	0,63
WATER	38	10	0,60
LIPIDS	3	103	0,49
LIVER	2	153	0,49
ERYTHROCYTE	13	23	0,48
CALCIUM	25	11	0,44
VENOUS	34	8	0,43
CAPILLARY	84	3	0,40
HEART	16	13	0,33
FISH	1	182	0,29
FATTY	1	178	0,28
ARTHRITIS	17	10	0,27
LUPUS	34	5	0,27
ATTACKS	71	2	0,23
PGI2	12	11	0,21
ANTIBODY	60	2	0,19
DIET	40	3	0,19
PULMONARY	59	2	0,19
TXB2	9	13	0,19
IMMUNOLOGIC	22	5	0,17
PROSTACYCLIN	22	5	0,17
TUMOR	6	18	0,17
HEALING	15	7	0,17
ERYTHEMATOSUS	34	3	0,16
SCLEROSIS	102	1	0,16
ULCERS	17	5	0,14
IMMUNE	28	3	0,13
PGF1	12	7	0,13
BLEEDING	4	19	0,12
PROTEIN	9	8	0,11
AUTOIMMUNE	14	5	0,11
RHEOLOGY	14	4	0,09
CANCER	7	7	0,08
CHOLESTEROL	1	45	0,07
ELECTRON	5	9	0,07
VESSEL	15	3	0,07
CIRCULATING	21	2	0,07
LUNG	7	6	0,07
MYOCARDIAL	10	4	0,06

ARACHIDONIC	1	37	0,06
DEFORMABILITY	9	4	0,06
CEREBRAL	7	5	0,06
ATHEROSCLEROSIS	8	4	0,05
HEMOSTATIC	5	6	0,05
RENAL	10	3	0,05
INFARCTION	9	3	0,04
IRON	3	9	0,04
TOCOPHERYL	3	9	0,04
FIBRINOLYSIS	13	2	0,04
IGG	13	2	0,04
NUTRITIONAL	5	4	0,03
CIRCULATORY	17	1	0,03
CHANNEL	16	1	0,03
HEMATOCRIT	4	4	0,03
THROMBOSIS	4	4	0,03
INFLAMMATORY	3	5	0,02
LEUKOCYTES	3	5	0,02
RHEUMATOID	15	1	0,02
DEGENERATION	5	3	0,02
INFECTION	5	3	0,02
HYPERPLASIA	2	7	0,02
NECROSIS	13	1	0,02
HYPERTENSIVE	6	2	0,02
KIDNEY	2	6	0,02
POLYMOR-PHONUCLEAR	3	4	0,02
CARDIAC	11	1	0,02
CLEARANCE	5	2	0,02
DIABETES	5	2	0,02
THROMBOGLOBULIN	10	1	0,02
URINE	5	2	0,02
CARDIOVASCULAR	3	3	0,01
COAGULATION	3	3	0,01
HYPERSENSITIVITY	9	1	0,01
PG	1	9	0,01
WOUND	3	3	0,01
ANTIPLATELET	4	2	0,01
BRAIN	2	4	0,01
HEPATIC	2	4	0,01
HORMONE	4	2	0,01
INDOMETHACIN	4	2	0,01
INTIMAL	1	8	0,01
PANCREATIC	4	2	0,01
AORTIC	2	3	0,01
ARTICULAR	6	1	0,01
CARBON	6	1	0,01
CORONARY	3	2	0,01

CYTOTOXIC	3	2	0,01
GLYCERYL	6	1	0,01
LIPOPROTEINS	1	6	0,01
OXYGEN	2	3	0,01
PEPTIDES	6	1	0,01
THROMBIN	1	6	0,01
TRINITRATE	6	1	0,01
ALBINO	1	5	0,01
FLUIDITY	1	5	0,01
URINARY	1	5	0,01
DIASTOLIC	5	1	0,01
MUSCULAR	5	1	0,01
AUTOLOGOUS	2	2	0,01
DYSTROPHY	4	1	0,01
MOUTH	2	2	0,01
SERINE	1	4	0,01
SODIUM	2	2	0,01
VASODILATORY	4	1	0,01
ALCOHOL	3	1	0,00
ANGINA	3	1	0,00
FEMORAL	3	1	0,00
MERCURY	1	3	0,00
CARBOHYDRATE	2	1	0,00
EYES	1	2	0,00
GLOMERULONE-PHRITIS	1	2	0,00
NEUROLOGICAL	2	1	0,00
NOREPINEPHRINE	2	1	0,00
THROMBOTIC	2	1	0,00
UREA	2	1	0,00
ABDOMINAL	1	1	0,00
AGING	1	1	0,00
ANGIOTENSIN	1	1	0,00
ANTIPLASMIN	1	1	0,00
ASPIRIN	1	1	0,00
CATECHOLAMINES	1	1	0,00
DIGESTIVE	1	1	0,00
EICOSANOID	1	1	0,00
KALLIKREIN	1	1	0,00
MEDITERRANEAN	1	1	0,00
NEUTROPHILIC	1	1	0,00
NORWEGIAN	1	1	0,00
NUTRITIVE	1	1	0,00
PROTHROMBIN	1	1	0,00
THROMBOPLASTIN	1	1	0,00

Tableau 19 : DPM sur les mots des titres et abstracts

Nous retrouvons les mêmes groupes de mots :

- *blood*, *platelet*, *artery*, *pressure*, *arterial*, *vascular*, *aggregation*, *viscosity*, *erythrocyte*, *venous*, *capilary*, *rheology*, *vessel*, *deformabiblity*,

hemostatic, *fibrinolysis*, *hematocrit*, *thrombosis*, *thromboglobulin*, *coagulation*, *antiplatelet*, *thrombin*, *fluidity*, *vasodilatory*, *thrombotic*, *antiplasmin*, *prothombin*, *thromboplastin*, liés, plus ou moins directement à l'hémorhéologie.

- des termes en rapport avec les prostaglandines et écosanoïdes, *prostaglandin*, *thromboxane*, *PGI2*, *TXB2*, *prostacyclin*, *PGF1*, *arachidonic*, *PG*, *eicosanoid*.

3.1.1.3 Remarques sur le travail sur les titres et/ou abstracts

Dans le cas de la recherche de liens entre l'huile de poisson et la maladie de Raynaud, bien que le travail soit plus simple en ne traitant que les titres et les abstracts les résultats obtenus avec les termes du MeSH sont plus faciles à exploiter. Le tableau issu du DPM réalisé avec les termes du MeSH (§ 2.4.4) est plus court, 20 lignes contre 59 ou 148 selon que nous nous intéressions aux titres seulement ou également aux abstracts. Les termes du MeSH que nous avons recherché sont des descripteurs relatifs à la physiologie, ce qui explique en partie le petit nombre de concepts communs aux deux littératures que retrouve le DPM. A l'inverse, aucun filtrage par type sémantique n'a été appliqué aux mots des titres ou des abstracts.

Les termes du MeSH sont des concepts, par définition, sans ambiguïté, alors qu'un mot du titre peut être interprété : si *thromboxane* est sans ambiguïté, *acid* doit être associé à d'autres mots pour gagner en signification, comme *arachidonic acid* par exemple. Lire un tableau de concept pour tenter d'en dégager quelques éventuelles tendances est plus rapide que d'effectuer le même travail sur un tableau de mots isolés. Parce que nous n'avons appliqué aucun filtre sémantique sur les titres et abstracts, la nature des concepts communs entre les deux littératures est variée : molécules, protéines, anatomie, physiologie, etc … Les termes du MeSH que nous avons recherchés dans notre première expérience DPM sont des descripteurs relatifs à la physiologie, ce qui explique en partie le petit nombre de concepts communs aux deux littératures que retrouve le DPM. Gordon et Lindsay ont reproduit le travail de Swanson sur la

maladie de Raynaud et l'huile de poisson [Gordon, 1996] en travaillant sur les bi-grammes et tri-grammes des textes des citations de Medline, plus pertinents que les mots isolés (voir § 1.7.3).

Notre technique de comptage présente le biais de calculer l'occurrence des mots et non leur fréquence. Ainsi, par exemple, en analysant titres et abstracts, *oil* apparaît 2 fois dans le corpus de la maladie de Raynaud, alors qu'une seule référence parle d'*oil* dans ce corpus, référence qui contient deux fois ce terme.

Le texte libre est soumis à la volonté des auteurs des articles, libre d'employer les termes avec lesquels ils sont les plus familiers. Les champs titres et abstracts sont "auteurs dépendants", alors que les termes du MeSH sont assignés à une notice suivant une charte de catalogage très détaillée et bien précise par des "indexeur" professionnels, ayant une compétence double : la connaissance d'une discipline biomédicale précise et la formation à l'indexation avec le MeSH. Cela n'empêche pas les indexeurs de commettre parfois quelques erreurs, corrigées par la NLM lorsqu'elles sont découvertes.

Pour clore ce paragraphe, le travail sur les titres et abstracts fait apparaître *oil* dans la colonne de la maladie de Raynaud. Il s'agit d'une référence que Swanson avait pointé *[Belch, 1985b]*, sur l'utilisation de l'EPO, *evening primerose oil* – huile essentielle de primevère – pour traiter le phénomène de Raynaud.

3.1.2 DPM et champs contrôlés

Les citations de Medline/PubMed comportent de nombreux champs dont le contenu est contrôlé. Outre les *MeSH Terms*, deux autres champs peuvent servir de support à une analyse DPM : *EC/RN Number* et *Secondary Source ID*.

3.1.2.1 EC/RN Number

Présent sous la forme [RN] dans la citation Medline, ce champ contient trois types de données, toutes contrôlés :

1. Les codes de *Chemical Abstracts Service* (CAS), les *Registry Numbers*, sous la forme : `RN  - 1553-41-9 (Eicosapentaenoic Acid)`
2. Les codes assignés par l'*Enzyme Commission* pour désigner une enzyme ou une protéine donnée ou une famille d'enzymes ou de protéines : `RN  - EC 2.7.1.37 (Protein Kinases)`
3. Des noms de molécules ou d'enzymes/protéines n'ayant pas encore de *Registry Number* ou de numéro *EC*, car non référencés par ces deux organismes :

```
RN  - 0 (Fish Oils)

RN  - 0 (LXRalpha protein)
```

Nous avons traité les littératures issues des équations précédemment décrites (§ 2.4.4 et § 3.1.1). Les codes numériques ont été éliminés afin de ne conserver que les noms des molécules ou protéines, plus explicites. Nous avons ensuite recherché les noms communs aux deux littératures, en présentant les résultats sous la forme de notre tableau DPM. Le coefficient est calculé sur la fréquence des noms.

RN/EC	Ray	Fish Oil	Coef.
Lipids	5	45	39,72
Epoprostenol	16	11	31,07
Prostaglandins	10	10	17,66
Cholesterol	4	19	13,42
Triglycerides	4	19	13,42
Prostaglandins E	12	6	12,71
Fatty Acids, Unsaturated	1	52	9,18
Dietary Fats	1	43	7,59
Arachidonic Acids	2	16	5,65
Ointments	4	7	4,94
Arachidonic Acid	2	12	4,24
Lipoproteins	2	12	4,24
Drug Combinations	3	7	3,71
Thromboxane B2	3	6	3,18
Phospholipids	1	17	3,00
Alprostadil	7	2	2,47
Vitamin E	1	14	2,47
Thromboxane A2	2	5	1,77
Collagen	3	3	1,59
6-Ketoprosta-glandin F1 alpha	4	2	1,41
Dinoprostone	2	4	1,41
Glucose	4	2	1,41
Norepinephrine	8	1	1,41
Placebos	8	1	1,41
Antigen-Antibody Complex	6	1	1,06
Calcium	2	3	1,06
Iron	2	3	1,06
beta-Thromboglobulin	4	1	0,71
Fatty Acids, Essential	1	4	0,71
Furans	4	1	0,71
Hemoglobins	2	2	0,71
Aspirin	3	1	0,53
Anti-Bacterial Agents	2	1	0,35
Anticoagulants	1	2	0,35
Delayed-Action Preparations	2	1	0,35
Epinephrine	2	1	0,35
Hydrocarbons	1	2	0,35
Indomethacin	2	1	0,35
Prednisolone	2	1	0,35
Serum Albumin	2	1	0,35
Theophylline	2	1	0,35
Alkaline Phosphatase	1	1	0,18
Alkenes	1	1	0,18
Antiplasmin	1	1	0,18
Barbiturates	1	1	0,18
Esters	1	1	0,18
Hydrocortisone	1	1	0,18
Immunosuppressive Agents	1	1	0,18
Insulin	1	1	0,18
Lactates	1	1	0,18
Potassium	1	1	0,18
Silicon Dioxide	1	1	0,18

Tableau 20 : DPM sur les champs RN

Le tableau 20 met en évidence les termes liés aux prostaglandines et aux eicosanoïdes : Epoprostenol, Prostaglandins, Prostaglandins E, Fatty Acids, Unsaturated, Arachidonic Acid, Thromboxane B2,

Alpostadil, Thromboxane A2, 6-Ketoprostaglandin F1 alpha, Dinoprostone. Les termes relatifs à la coagulation sont en nombre plus restreint et plus bas dans le tableau : beta-Thromboglobulin, Anticoagulants, Antiplasmin. Ces résultats mettent nettement en avant les prostaglandines et analogues comme lien possible entre l'huile de poisson et la maladie de Raynaud, incitant le biologiste à travailler dans ce sens.

3.1.2.2 Secondary Source ID

Le champ *SI* identifie des sources d'informations secondaires, ainsi que les numéros d'accès uniques aux références de ces sources, mentionnées dans les citations Medline. Ce champ est composé du nom de la source citée suivi du numéro d'accès à la référence :

```
SI  - GENBANK/A34936
SI  - SWISSPROT/P83574
SI  - OMIM/121800
```

Plusieurs valeurs provenant de sources diverses peuvent le composer :

- Genbank[38], base de données de séquences génétiques du NIH, collection annotée de séquences d'ADN publiques. Fin 2004, GenBank contenait plus de 40 millions de séquences totalisant 44,5 milliards de nucléotides.
- OMIM[39], *Online Mendelian Inheritance in Man*, catalogue de gènes humains et de maladies génétiques, créé et édité par le Dr. McKusick de John Hopkins et ses collègues. OMIM a été développé pour l'Internet par le NCBI et propose les fiches des maladies et gènes ainsi que de nombreux liens vers Medline ou vers d'autres bases de données du système *Entrez*.

[38] NCBI. (Page consultée le 22 septembre 2005). *GenBank Overview*, [En ligne]. Adresse URL : http://www.ncbi.nlm.nih.gov/Genbank/

[39] NCBI/John Hopkins University. (Page consultée le 22 septembre 2005). *OMIM, Online Inheritance in Man*, [En ligne]. Adresse URL : http ://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

- ClinicalTrials.gov[40], répertoire d'essais cliniques sponsorisés par des institutions américaines ou des entreprises, qui se déroulent aux Etats-Unis.
- Swiss-Prot[41], base de données de protéines. Chaque entrée comprend des annotations (organisme, synonymes, codes, etc…) des références bibliographiques ainsi que des liens vers d'autres sources d'informations.

D'autres sources alimentent le contenu de ce champ.

Les corpus générés par les équations sur l'huile de poisson et sur la maladie de Raynaud (§ 2.4.4 et § 3.1.1) ne contiennent aucun champ *SI*, certainement parce qu'en 1985 la NLM n'avait pas encore créé *SI*. Les citations disponibles sur PubMed le 26 août 2005, citations relatives à la maladie de Raynaud d'une part et, d'autre part à l'huile de poisson, contiennent bien pour certaines des éléments renseignant le champ *SI*. Cependant les deux littératures n'ont pas d'éléments *SI* communs.

SI pourrait être employé pour mettre en évidence des protéines ou des séquences (nucléiques ou protéiques) communes à deux corpus. Il pourrait s'agir d'une maladie ou d'un effet physiologique et d'une molécule, pour tenter d'identifier un mécanisme d'action potentiel. Un lien entre deux maladies pourrait également être ainsi découvert.

[40] National Institutes of Health/National Library of Medicine. (Page consultée le 22 septembre 2005). *ClinicalTrials.gov*, [En ligne]. Adresse URL : http:// clinicaltrials.gov

[41] Swiss Institute of Bioinformatics/EMBL Outstation - The European Bioinformatics Institute (EBI). (Page consultée le 22 septembre 2005). *Swiss-Prot Protein knowledgebase - TrEMBL Computer-annotated supplement to Swiss-Prot*, [En ligne]. Adresse URL : http://ca.expasy.org/sprot

3.2 Découverte de connaissances et autres sources d'information

Habituellement, les recherches et travaux de découverte de connaissances et de génération d'hypothèse dans le domaine biomédical s'effectuent en utilisant la "matière première" disponible à travers PubMed. Nous voyons trois raisons à cela. Tout d'abord PubMed offre un accès libre à Medline, or les techniques de KDD traitent bien souvent de très grandes quantités de citations. Il nous est arrivé de travailler sur des volumes équivalents ou dépassant les 100.000 citations. Ensuite, PubMed est une interface offrant toute une série de fonctionnalités documentaires très évoluées – à notre avis supérieures à bien des serveurs professionnels – qui facilitent l'exploitation des données (interface de recherche simple et très complète, thesaurus en ligne, historique des équations de recherche évolué, connexions et liens avec d'autres bases du système *Entrez* (OMIM, GenBank, Proteins) etc … Enfin, la NLM met gratuitement à disposition de tous les outils qu'elle a développé. C'est le cas du MeSH et d'UMLS. Cet ensemble unique offre de multiples possibilités : par exemple, à partir d'une maladie, chercher les liens qui aboutissent à une origine génétique [Weeber, 2001].

Cependant, d'autres sources d'informations peuvent être exploitées : bases de données bibliographiques ou l'Internet.

3.2.1 Bases de données bibliographiques

L'information contenue dans Embase est structurée d'une manière similaire à celle contenue dans Medline. En particulier, Embase s'appuie sur un thesaurus, le EMTREE[42], pour indexer les citations. Ce thesaurus offre une structure hiérarchique et une classification qui permet d'isoler et de créer des listes de concepts relatifs aux molécules, aux phénomènes physiologiques ou à l'anatomie. Sur le principe, l'utilisation d'Embase et du EMTREE pour réaliser un DPM est tout à fait envisageable. La principale difficulté à contourner n'est pas l'accès gratuit aux citations puisque le serveur DataStar – par exemple –

[42] Elsevier B.V. (Page consultée le 22 septembre 2005). *Embase suite of products*, [En ligne]. Adresse URL : http://www.info.embase.com/emtree/

intègre les descripteurs du EMTREE dans le format gratuit de visualisation des citations d'Embase dans la limite de 2.000 par déchargement, mais cette difficulté réside dans l'accès à l'EMTREE qui, à notre connaissance, n'est pas disponible librement.

Rappelons qu'Embase est intégrée au DPM puisque lors de l'analyse des tableaux, de nombreuses bibliographies sont réalisées sur Medline, mais également Embase, Pascal, Biosis ou les Derwent Drug Files.

3.2.2 Découverte de connaissances et Internet

Internet, le *World Wide Web*, offre un formidable espace d'exploration pour tenter de mettre en lumière des connaissances cachées. En janvier 2005, il était estimé que la taille du *web* était alors de 11,5 milliards de pages [Gulli, 2005]. Cette estimation rend compte du nombre de pages *web* indexables par quatre moteurs de recherche majeurs : Google, Yahoo!, Ask/Teoma et MSN. Ce volume considérable d'informations et le développement des connexions à haut débit font que le *Web* est la première source vers laquelle on se tourne lors d'une recherche d'informations. Cependant, son hétérogénéité, tant en terme de formats, de modes d'accès ou de pratiques rédactionnelles rend le *Web* de plus en plus difficilement accessible aux traitements automatisés [Cadel, 2005]. Ainsi les contenus disponibles par l'intermédiaire de formulaires d'interrogations et stockés dans des bases de données nécessitent de développer des applications spécifiques pour y accéder automatiquement. Les scripts sont aussi des entraves à l'automatisation de la recherche d'informations. Or, dans les techniques de découvertes de connaissances font appel à l'exploitation massive d'informations impossibles à réaliser manuellement ; l'automatisation de la récupération des données est indispensable.

La découverte de connaissances sur Internet se trouve confronté au problème du vocabulaire. Dans leur grande majorité les pages *web* ne sont pas indexées comme le sont les citations des bases de données bibliographiques. Ce qui impose d'employer des outils de langage naturel ou qui assurent la conversion des mots en concepts ou descripteurs prédéfinis afin de rendre leurs contenus

plus facilement manipulables. Travailler sur les descripteurs épargne les ambiguïtés et les difficultés que l'on rencontre sur le texte intégral, telles que synonymie, polysémie ou acronymes. Loh et coll. [Loh, 2000] ont élaboré un système de KDT (*Knowledge Discovery in Texts*) basé sur les l'extractions de concepts de textes issus de l'Internet. L'avantage des concepts est d'offrir à l'utilisateur d'un tel système de saisir rapidement le sens des textes analysés et d'en résumer rapidement les éléments centraux. Cette approche permet de déterminer des tendances d'associations de concepts. Les usages de ce type de méthode sont multiples : analyse de discours, sociologie, recherche biomédicale (recherche de relation entre différents symptômes) et veille technologique (stratégies développées par différents compétiteurs). Cependant, il ne s'agit pas à proprement parler de génération d'hypothèse.

Gordon et Lindsay, qui avaient simulé les expériences de Swanson à partir d'une méthode basée sur des statistiques lexicales (voir § 1.7.3) [Gordon, 1996] et [Lindsay, 1999], ont également travaillé sur l'utilisation d'Internet dans un système de découverte de connaissances basé sur la littérature [Gordon, 2002]. Ils proposent de généraliser le modèle ABC, centré sur la découverte de traitements pour une maladie donnée (ou ses dérivés),

molécule existante → profil pharmacologique/effets secondaires →

nouvelle application

en un modèle ouvert pour la découverte de connaissances au sens large :

point de départ → littérature intermédiaire → découverte

Cela pourrait être, par exemple :

technologie existante → caractéristiques majeurs de cette technologie →

nouvelle application

Gordon et Lindsay expliquent également que les techniques de découverte de connaissances peuvent aider une personne donnée à appréhender un domaine ou un problème qui ne lui est pas familier. La découverte de connaissances n'est pas limitée à la recherche de nouveautés dans l'absolu, mais elle est aussi un outil d'exploration à la disposition d'un individu soucieux d'augmenter son

savoir par l'exploration d'une thématique qu'il ne connaît pas. Il s'agit dans ce cas de rechercher et mettre en évidence les éléments communs entre sa connaissance et la thématique qu'il explore. Un système qui permet à un individu d'apprendre et d'explorer les relations qui existent entre ses connaissances et un problème dont la solution lui est inconnue est à ce titre un système de découverte de connaissances. Par exemple, un directeur des ventes qui souhaite trouver des cas de succès de lancement de produits dans d'autres secteurs industriels que le sien (secteurs présentant cependant des liens entre eux) est intéressé par une information nouvelle de son point de vue. Peu importe que d'autres personnes connaissent cette information. La découverte de connaissances basée sur la littérature est une technique pertinente dans un tel cas.

Gordon et Lindsay présentent différentes approches de découvertes de connaissances basées sur l'Internet, autour d'un même thème, les algorithmes génétiques[43].

La première expérience qui utilise les statistiques lexicales, a pour objectif de trouver de nouvelles applications à des techniques – solutions – existantes. Il s'agit d'étendre ces solutions à d'autres domaines que ceux pour lesquels elles ont été initialement conçues. Dans le cas des algorithmes génétiques, les auteurs interrogent AltaVista[44] avec la requête *genetic algorithms* et récupèrent le contenu des 50 documents les plus importants en les traitant avec MemoWeb[45], un outil de capture de contenu de pages *web*. Les termes composés de deux mots (bi-grammes) sont isolés et leurs fréquences sont calculées afin d'établir les statistiques lexicales. Douze termes en relation avec les algorithmes génétiques sont sélectionnés parmi 3.000 en exploitant les statistiques. Ces termes servent à leur tour de requête pour interroger AltaVista. Les différents traitements

[43] Les algorithmes génétiques ont pour but d'obtenir une solution approchée, en un temps acceptable, à un problème d'optimisation lorsqu'il n'existe pas de méthode exacte pour le résoudre. Ils utilisent la notion de sélection naturelle développée par Darwin et l'appliquent à une population de solutions au problème donné (*d'après* Wikipédia. (Page consultée le 22 septembre 2005). *Wikipédia*, [En ligne]. Adresse URL : http://fr.wikipedia.org).

[44] Overture Services, Inc. (Page consultée le 22 septembre 2005). *Altavista*, [En ligne], Adresse URL : http://www.altavista.com

appliqués aux 100 premiers documents correspondants à chacune des douze requêtes mettent en avant 42 termes (bi-grammes), dans une liste de 8.000, dont chacun est une découverte potentielle de nouvelle application des algorithmes génétiques. Par exemple, Gordon et Lindsay proposent qu'un algorithme génétique soit employé dans un modèle financier de simulation de portfolio optimisé en terme de risque et retour sur investissement.

Leur deuxième expérience montre que leur méthode de découverte de connaissances peut être employée pour stimuler l'inspiration. Des concepts connexes aux algorithmes génétiques sont identifiés par une personne, concepts qui sont porteurs de sens et peuvent potentiellement conduire à de nouvelles découvertes. Quatre concepts sont ainsi identifiés, intellectuellement, et sont utilisés pour interroger AltaVista : *search algorithms* ou *search techniques*, *optimization*, *machine learning* et *adaptative algorithms* ou *adaptative searches*. Les documents recueillis sur le *web* proviennent de requêtes employant les quatre concepts et éliminant *genetic algorithms* (par exemple *search techniques* NOT *genetic algorithms*). Quatre littératures sont analysées pour extraire les termes de deux mots. Selon la littérature, les listes générées varient de 3.000 à 10.000 termes, qui sont triés selon leurs fréquences dans les documents. Les auteurs s'intéressent ensuite aux termes de fréquences faibles, mais supérieures ou égales à deux et identifient "à la main", pour chaque liste, les dix termes qu'ils jugent les plus prometteurs. Il s'agit de rechercher la nouveauté : un terme de faible fréquence peut être un concept moins étudié et par conséquence, le lien entre ce concept et les algorithmes génétiques sera d'autant plus ténu. L'analyse des quatre littératures produit une liste de 37 concepts, qui sont croisés avec algorithmes génétiques dans une recherche sur Web of Science[46] et UseNet[47] pour estimer les liens directs entre chaque terme et algorithmes génétiques : six des 37 concepts n'ont aucun lien avec algorithmes génétiques. A ce stade, un

[45] GOTO Software. (Page consultée le 22 septembre 2005). *Memoweb*, [En ligne]. Adresse URL : http://www.goto.fr/memoweb/index.asp

[46] Thomson. (Page consultée le 22 septembre 2005). *Web of Science*, [En ligne]. Adresse URL : http://www.isinet.com/products/citation/wos

[47] Usenet.com. (Page consultée le 22 septembre 2005). *Usenet.com*, [En ligne]. Adresse URL : http://www.usenet.com

expert peut intervenir pour estimer l'opportunité d'exploiter les algorithmes génétiques dans l'une ou l'autre des thématiques suggérées par ces concepts.

Dans leur troisième expérience, Gordon et Lindsay ont voulu répliquer l'expérience précédente, sans intervention humaine : ils ne sont pas arrivés à automatiser le processus de découverte. L'intervention humaine est nécessaire, en particulier pour la sélection des dix concepts intermédiaires en utilisant les termes de faibles fréquences. Ces termes étant les plus nombreux, il n'est pas possible d'automatiser cette tache.

Enfin, leur dernière expérience a consisté à solliciter directement John Holland, "père" des algorithmes génétiques, en lui demandant d'identifier de nouvelles applications. Holland formula quatre propositions dont aucune ne figurent dans les tableaux générés par Gordon et Lindsay. La conclusion de cette expérience est que même si les hypothèses générées par Gordon et Lindsay peuvent individuellement être également formulée par tel ou tel chercheur ou expert, aucun individu n'est près d'identifier une telle quantité d'hypothèses différentes.

Pour conclure, Gordon et Lindsay ont montré différentes applications des méthodes de découverte de connaissances basées dans la littérature :

- identifier de nouvelles applications pour la solution à un problème donné,
- exploiter leur méthode d'analyse lexicale pour aider un chercheur à dégager de nouvelles idées ou pistes de travail en relation avec sa problématique.

Notons que l'analyse d'un expert et l'intervention humaine demeure les points clés de la réussite des différentes expériences présentées dans cet article.

3.3 Conclusion de la troisième partie

Le DPM s'appuie sur le MeSH pour extraire des concepts de la littérature. Ces concepts sont ensuite traités de manière non booléenne et présentés sous forme de tableaux à n colonnes. Les exemples précédents nous montrent que le DPM peut être adapté pour fonctionner sur d'autres types d'informations que les termes du MeSH, comme le texte libre (mots du titre et de l'abstract) ou d'autres champs codés (*RN* et *SI*).

Une voie que nous n'avons pas explorée, mais prometteuse à nos yeux est la recherche de combinaisons de médicaments pour traiter une maladie. C'est l'exemple connu de la HAART - *Highly Active AntiRetroviral Therapy* – combinaison de plusieurs antirétroviraux pour traiter les personnes atteintes par le virus de l'immunodéficience humaine. Pour une maladie donnée, procédant de manière itérative, le DPM pourrait servir à proposer la meilleure hypothèse de combinaison de molécules, agissant de manière efficace sur l'ensemble des dérèglements physiopathologiques.

Les travaux en KDD ont montré qu'il est possible d'exploiter pratiquement n'importe quelle source textuelle, à condition d'en rendre le contenu accessible à un ordinateur. La difficulté de telles opérations est l'identification et l'extraction des concepts pertinents contenus dans un texte afin de pouvoir les traiter statistiquement. Pour clore cette partie, mentionnons que les brevets constituent une source très importante d'information, utilisable en KDD [Markellos, 2002] et [Mukherjea, 2005].

A discovery is said to be an accident meeting a prepared mind.

Albert von Szent-Gyorgyi

CONCLUSION

4.1 DPM, industrie pharmaceutique et expertise

Les outils du DPM ne sont rien s'ils ne sont pas employés dans le cadre d'une double expertise pour tenter de répondre à une question. Expertise biomédicale d'une part et, d'autre part, en traitement de l'information. La question, point de départ de la réflexion va servir de fil rouge aux deux expertises pour conduire le DPM, dans le but d'apporter une réponse sous la forme d'une hypothèse :

- expertise biomédicale : analyse du problème, formulation de la question, recherche des liens biologiques possibles entre les éléments disjoints qui font l'objet du DPM, analyse des résultats et de la littérature, élaboration et test d'hypothèses … Cette expertise s'appuie sur une bonne connaissance du sujet traité, mais également des connaissances globales en physiologie, pharmacologie et en médecine. Ajoutons que l'ouverture d'esprit et la curiosité scientifique sont deux qualités indispensables.
- expertise en traitement de l'information : analyse du problème, formulation de la question, traduction des concepts en descripteurs, réalisation des opérations techniques du DPM, recherches bibliographiques … Ouverture d'esprit et curiosité scientifique sont ici aussi deux qualités indispensables, tout comme une bonne connaissance du domaine biomédical.

Le DPM, comme tout traitement bibliométrique ou de *text-mining* n'a de sens que s'il a pour but de répondre à une question en fournissant des éléments à

l'expert [Kostoff, 2001]. Ces éléments doivent aboutir à une prise de décision. Dans notre cas, cette décision peut être l'initiation d'un programme de recherche fondamentale ou clinique, la réalisation d'expériences ponctuelles pour confirmation d'hypothèses, la recherche d'entreprises pour un partenariat, etc … Le DPM apporte de l'information à l'expert afin qu'il formule des recommandations en vue de décisions. Dans un environnement de concurrence toujours plus forte, la prise de décision est un acte managérial fondamental [Galland, 2004, 2005]. Selon Galland, "*il existe principalement trois typologies de décisions :*

1. *Les décisions stratégiques : les décisions stratégiques concernent les relations de l'entreprise avec son environnement et conditionnent essentiellement la manière dont l'entreprise va se positionner sur un marché.*
2. *Les décisions tactiques : les décisions tactiques doivent permettre de définir et d'organiser les ressources de l'entreprise dans la réalisation des objectifs fixés dans le cadre des décisions stratégiques.*
3. *Les décisions opérationnelles : les décisions opérationnelles portent sur l'exploitation courante de l'entreprise et concerne l'utilisation optimale des ressources allouées au processus productif de l'entreprise.* " [Galland, 2005].

Dans le cadre de l'industrie pharmaceutique, le DPM est positionné comme un support d'aide à la décision stratégique étant donnés les coûts et durées relatifs au développement d'un produit. En effet, une dizaine d'années sont nécessaires pour mettre un nouveau médicament sur le marché, avec un taux d'échec considérable, pour un coût moyen de développement de 802 millions de dollars US [Lawrence, 2002], [DiMasi, 2003]. Ce chiffre est à mettre en perspective avec les données mondiales 2004 d'IMS[48] : le chiffre d'affaire dégagé par

[48] IMS, Intercontinental Marketing Service. (Page consultée le 20 janvier 2006). *IMSHealth*, [En ligne]. Adresse URL : http://www.imshealth.com

l'industrie pharmaceutique est d'environ 530 milliards de dollars US dont 270 ont été réalisés par les 11 premiers laboratoires. En moyenne, les 50 premiers laboratoires pharmaceutiques ont réalisés chacun 8,2 milliards de dollars. Il est admis qu'en 2003 l'industrie pharmaceutique a consacré 18% de son chiffre d'affaire à l'activité de recherche et développement [Meek, 2004]. DiMasi considère qu'un médicament approuvé aux Etats-Unis entre 1990 et 1994 rapportera en moyenne 2,4 milliards de dollars tout au long de son cycle de vie ; cette estimation se base sur l'analyse de quatre classes de produits (analgésiques/anesthésiques, anti-infectieux, cardiovasculaires et médicaments du système nerveux central) [DiMasi, 2004].

Classiquement, le processus de recherche et développement est divisé en six phases successives [Warne, 2003] :

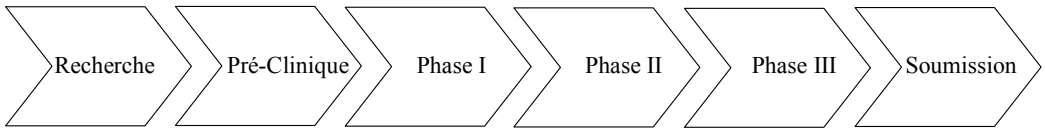


Figure 16 : Phases de recherche et développement d'un médicament

La recherche a pour objet d'identifier une molécule active sur une pathologie donnée. La phase pré-clinique regroupe tout un ensemble d'activités : formulation du principe actif, études toxicologiques, travail de mise à l'échelle pour la production de la molécule, études de pharmacocinétique et métabolisme. C'est à partir de cette phase que commence à se constituer le dossier d'enregistrement du futur médicament. La phase I doit démontrer, chez le volontaire sain, que la molécule est bien tolérée à la dose à laquelle elle est active. L'efficacité et l'innocuité de la molécule sont testées sur de petits nombres de patients en phase II afin d'estimer le rapport bénéfice/risque. Enfin, la phase III doit confirmer à large échelle les résultats des phases II. L'ultime

étape est la soumission du dossier d'enregistrement auprès des autorités réglementaires des pays – ou zones géographiques – où la molécule devenue médicament doit être commercialisée. Il s'agit d'une vision résumée et simplifiée du processus, où les étapes ne suivent pas de manières strictes, certaines pouvant être conduites de front et où l'activité inventive n'est pas limitée à la recherche. Les informations recueillies dans les phases en aval du processus peuvent être reprises dans des phases plus précoces pour optimiser le développement. Habituellement, le développement englobe les étapes allant de la pré-clinique à la soumission.

La recherche est elle-même divisée en six phases interactives [Warne, 2003] :

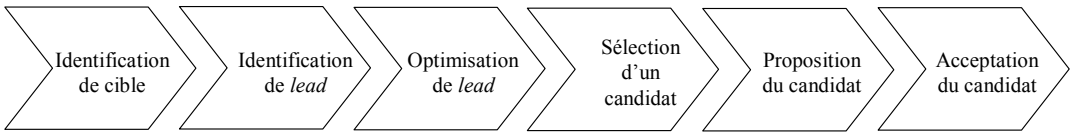
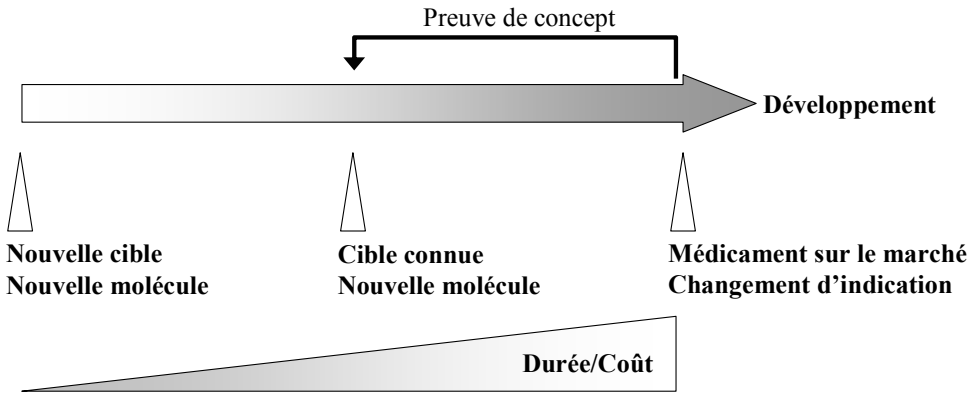


Figure 17 : Phases de recherche d'un médicament

La cible est classiquement un gène ou une protéine dont on espère que l'usage ou la modification va produire un effet thérapeutique pour une pathologie donnée. Le *lead* désigne une molécule qui a une activité sur la cible. L'identification permet de découvrir quels types de molécules peuvent agir sur la cible, l'optimisation sélectionne les *leads* les plus actifs pour le soumettre au processus de sélection d'un candidat qui sera proposé pour un développement.

Le DPM peut intervenir au niveau des décisions stratégiques de la phase de recherche, mais également en développement. Nous proposons ici quatre applications principales suivant les stratégies de recherches de molécules :

1. Identification de nouvelles cibles et de nouvelles molécules.
2. Identification de molécules pour une cible connue.
3. Changement d'indication pour un médicament déjà sur le marché.
4. Une variante de la troisième application est de chercher de nouveaux analogues d'un médicament déjà sur le marché dans une autre indication que celle pour laquelle il a été enregistré. Il est ainsi possible d'utiliser ce médicament à des fins expérimentales pour prouver que ce concept est valable.



D'après Fabrizio Dolfi

Figure 18 : Stratégies de recherche et développement d'un médicament

Trouver une nouvelle cible et développer une nouvelle molécule coûte plus et dure plus longtemps que de réussir à changer l'indication d'un médicament enregistré. L'emploie du DAD pour proposer de nouveaux usages de la thalidomide correspond au changement d'indication [Weeber, 2003]. Le DPM réalise le même type d'opération : pour un médicament donné, quels sont ses actions pharmacologiques connues (ou bien celles de la classe à laquelle il

appartient) ? A partir de là, le DPM va aider à identifier quelles pathologies sont concernées par ces actions pharmacologiques.

Le DPM peut donc intervenir en plusieurs points critiques du processus de recherche et développement d'un médicament, points où la prise de décision est stratégique et ou l'analyse de l'information par une expertise éclairée est indispensable. Classiquement, pour répondre à un besoin de connaissance, le cycle de l'information commence par l'identification des sources pertinentes, sa collecte, son évaluation, son analyse et son traitement, puis sa synthèse. C'est un processus itératif. Le DPM suit ce cycle, facilitant les étapes de traitement afin de permettre à l'expert de se concentrer sur l'analyse et la synthèse afin de produire de solides hypothèses.

4.2 Diffusion du modèle de Swanson

Nous avons montré qu'il existe un certain nombre de travaux inspirés du modèle et/ou des exemples de Don Swanson (une quinzaine d'articles dans les sciences du vivant à notre connaissance). Pratiquement tous les travaux de génération d'hypothèse dans le domaine biomédical sont basé sur la mise en lumière de structures complémentaires et disjointes par l'exploitation de la littérature [Cohen, 2005]. La croissance du volume de données, qu'elles soient bibliographiques, génomiques ou protéomiques nous laisse imaginer que d'importantes découvertes sont masquées par l'explosion de la quantité d'information. Cependant, les systèmes de KDD ne font pas encore partie des outils standards des biologistes. Cohen avance le fait que ces outils doivent être encore améliorés, tant en terme de performance (volume de données, information de natures différentes) que d'évaluation des résultats produits.

Le travail de Swanson fut accueilli de différentes manières, selon que l'on se place du côté des sciences de l'information ou du côtés des sciences du vivant. Spasser a publié une analyse des citations des travaux de Swanson afin de suivre comment ses idées ont été perçues par la communauté scientifique [Spasser,

1997]. Il a ainsi relevé 21 articles en sciences de l'information et 12 en sciences de la vie.

En sciences de l'information, les auteurs considèrent la méthode et les techniques employées, les hypothèses médicales générées étant pour eux des illustrations de sa méthode. Ils ne discutent pas l'existence du savoir public caché, c'est pour eux un fait établi. Tout comme la possibilité de générer de nouvelles connaissances par l'usage ingénieux de systèmes d'information. Le savoir implicite soutenu par des informations disjointe est une réalité – que tout professionnel de l'information pratiquant la recherche non booléenne connaît. Certains auteurs mettent en avant le rôle du professionnel de l'information spécialisé dans le domaine biomédical et le qualifient de fondamental dans la création de nouvelles connaissances [Schell, 1992]. Le documentaliste biomédical – *medical librarian* – est dans une position unique pour identifier les connections logiques entre des littératures non interactives. Ils peuvent jouer un rôle dynamique dans l'exploration de la littérature et donc influencer l'orientation des recherches dans le cadre de la découverte de connaissance.

Dans les sciences de la vie, Spasser montre que, de manière attendue, les auteurs s'intéressent aux hypothèses formulées plutôt qu'à la méthode. Leurs discours à l'égard des hypothèses est plutôt négatif et condescendant, bien qu'ils s'en servent pour étayer leurs argumentaires. Dans ce cas, les hypothèses de Swanson sont en attente de vraies validations scientifiques. Selon Spasser les normes de la recherche biomédicale traditionnelle sont en désaccord avec la recherche basée sur la littérature proposé par Swanson. Le but de la recherche scientifique est de prédire et de contrôler, pas de comprendre le fruit de découvertes inattendues ; elle est orientée vers la vérification plutôt que vers la découverte ; son approche est concentrée sur les particularités au lieu d'être ouverte sur une appréhension globale des problèmes ou des questions. Spasser termine ainsi : "*... finally, its ethos is originality and priority, not synthesis and bridge-building*". Il ne voit rien d'étonnant à ce que les chercheurs des sciences de la vie soient peu disposés à considérer une méthode exploratoire qui ne peut être testée de manière empirique ou quantitative.

Dans quelques cas, les travaux de Swanson ne font pas l'objet d'une rhétorique négative : les auteurs utilisent alors les articles de Swanson pour mentionner une relation ou un fait établi de manière empirique.

En conclusion de son papier Spasser souligne que bien qu'utile à la recherche biomédicale et particulièrement intéressant, le concept de savoir public caché ainsi que la méthode pour le mettre en lumière demeurent largement inconnu des chercheurs. En conséquence, il représente lui-même un savoir public caché.

Depuis 1997, date à laquelle fut publié l'article de Spasser, quelques papiers associant spécialistes des "deux bords" - sciences de l'information et sciences du vivant – ont été publiés[49] : outre Swanson et Smalheiser, citons Weeber et Molema [Weeber, 2003], Persidis et Deftereos [Persidis, 2004], Wren et Bekeredjian [Wren, 2004b], sans oublier notre propre contribution [Pierret, 2005].

Pour l'anecdote, un papier paru en 2004 mentionne l'article de Spasser [Spasser, 1997], le décrivant comme une analyse de l'exploitation des littératures disjointes afin d'en faire émerger le savoir public caché, sans mentionner le nom de Swanson dans ce contexte [Abbott, 2004]

[49] Par simplification pour chaque article cité, nous mentionnons le premier auteur en science l'information, puis le premier auteur en sciences du vivant.

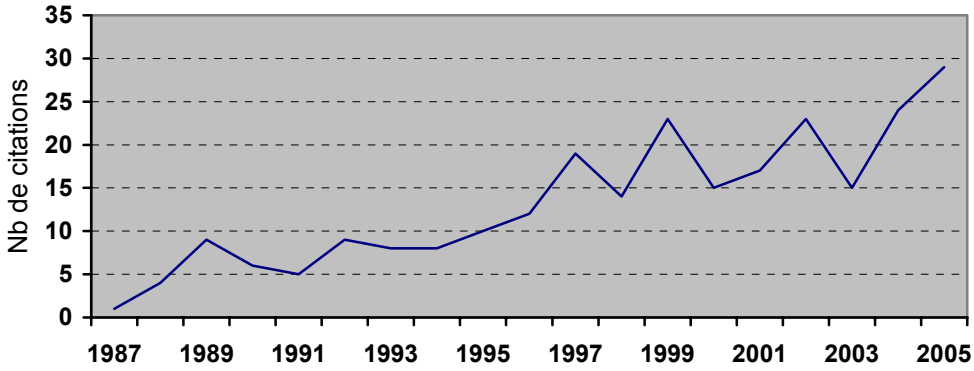


Figure 19 : Suivi des citations des articles publiés par Swanson sur sa méthode ou sur ses hypothèses

La figure 19 illustre l'évolution des citations au cours du temps des articles publiés par Don Swanson sur sa méthode ou sur ses hypothèses. Il a été réalisé le 21 septembre 2005 a partir de l'interrogation de SciSearch et de Social SciSearch sur le serveur DataStar, en éliminant les autocitations provenant de Swanson ou de Smalheiser. La tendance générale est que ses écrits suscitent de plus en plus l'attention des chercheurs.

L'interrogation a été conduite à partir des citations des 16 articles suivants : [Swanson, 1990a], [Swanson, 1990b], [Swanson, 1989b], [Swanson, 1989a], [Swanson, 1988], [Swanson, 1986a], [Swanson, 1987], [Swanson, 1986b], [Swanson, 1997], [Swanson, 1999], [Swanson, 2001b], [Smalheiser, 1998a], [Smalheiser, 1998b], [Smalheiser, 1996b], [Smalheiser, 1996a] et [Smalheiser, 1994].

4.3 Retour sur le travail de Swanson

Don Swanson a largement détaillé le mode opératoire de sa découverte, basé sur une utilisation classique de bases de données bibliographiques (Medline ou SciSearch) et les croisements méthodiques et systématiques effectués ensuite : il n'existe aucun lien entre maladie de Raynaud et huile de poisson avant

novembre 1985. Depuis, une quinzaine d'articles ont été publiés sur ce sujet, dont le quart parlent directement du travail de Swanson. Ce travail a été répliqué à plusieurs reprises, selon des modalités différentes [Gordon, 1996], [Swanson, 1999], [Weeber, 2000], [Srinivasan, 2004]. Les résultats obtenus permettent toujours de proposer l'utilisation de l'huile de poisson comme traitement. L'hypothèse qu'il a proposée n'était pas si évidente qu'il n'y parait à posteriori – sinon, d'autres l'auraient fait avant lui, des cliniciens par exemple. En revanche, le raisonnement qui conduit à la formuler nous apparaît comme limpide : si une substance A agit sur un ou plusieurs phénomènes physiologiques B, impliqués dans une maladie C, alors A peut potentiellement être un traitement de C. A partir de ce raisonnement, par transitivité, il est facile d'intégrer les éléments qui conduisent à proposer son hypothèse :

- la maladie de Raynaud est caractérisée, entre autres, par des problèmes d'agrégation plaquettaire et une augmentation de la viscosité sanguine,
- l'huile de poisson inhibe l'agrégation plaquettaire et diminue la viscosité sanguine,
- il est donc probable que l'huile de poisson améliore l'état des personnes atteintes de maladie de Raynaud.

Il s'agit d'un savoir latent, qui est resté caché de la communauté scientifique tant que le rapprochement huile de poisson/maladie de Raynaud n'a pas été révélé par Swanson. Son travail est devenu un exemple pour ceux qui ont par la suite tenté d'automatiser le processus de découverte ; il permet de tester différentes méthodes et de vérifier si elles donnent les éléments nécessaires à la formulation de l'hypothèse. Le point commun entre tous ces travaux autour de la première découverte de Swanson est l'intervention d'experts du domaine médical : quelle que soit la technique utilisée (UMLS, MeSH, mots du texte, DPM …), l'expert joue un rôle central dans le pilotage du système, en identifiant ce qui est important ou insolite et en gardant une bonne dose d'imagination pour ne pas se couper d'opportunités de nouvelles découvertes. Swanson et Smalheiser insistent bien sur ce point à propos de l'utilisateur d'Arrowsmith : les utilisateurs doivent savoir exploiter les bases de données bibliographiques et les thesaurus associés

avant toute utilisation du logiciel. Ils doivent aussi bien connaître le sujet sur lequel ils travaillent, faire preuve d'ingéniosité et être capables de repérer des connections prometteuses [Swanson, 1999]. Il en est de même pour les autres méthodes. Un tel travail s'accompagne de nombreuses interrogations directes des bases de données bibliographiques afin d'identifier les littératures complémentaires, d'une part et, d'autre part, de s'assurer que l'hypothèse que l'on tente de formuler n'est pas connue. A titre indicatif, le DPM permet de générer rapidement la liste des termes B. Il repose sur la collaboration d'un spécialiste de l'information biomédicale et d'un expert en physiologie. Pour l'exemple de la maladie de Raynaud, il faut un peu plus d'une heure pour générer les termes des premières requêtes et produire les premières listes. L'analyse des résultats est plus longue et implique principalement l'expert, mais aussi le spécialiste de l'information qui intervient ponctuellement pour réaliser des bibliographies ciblées ou générer de nouvelles listes à la demande de l'expert. Deux jours ont été nécessaires pour formuler l'hypothèse de l'adénosine et nous estimons qu'il faudrait trois à cinq autres jours nécessaires pour faire le point sur l'ensemble des éléments bibliographiques disponibles sur le sujet.

Cette durée peut paraître longue si on se place dans le contexte du travail de recherche bibliographique traditionnel, puisque de l'ordre de la semaine. Elle est en revanche sans commune mesure avec la durée de la phase de recherche dans l'industrie pharmaceutique, qui varie de 2 à 5 ans. Aujourd'hui, la recherche pharmaceutique se dote de techniques de découverte *in silico* (modélisation moléculaire, screening virtuel, outils bioinformatiques) et génère de plus en plus de données par des technologies à haut débit afin de réduire la durée de la phase exploratoire et des phases précoces de la recherche. La "fouille du bibliome" entre maintenant dans son arsenal (voir par exemple les outils développés par le NCBI). Le knowledge discovery in databases n'est pas loin.

Quel que soit l'outil utilisé – Arrowsmith, le DAD, le DPM – le schéma de Swanson et les techniques associées sont des aides précieuses à l'exploitation

des données bibliographiques. PubMed contenait le 15 septembre 2005 15.806.106 citations, soit 464.371 de plus que le 23 janvier dernier, date à laquelle l'introduction de ce mémoire a été rédigée. La quantité de paires possibles d'articles de Medline a elle augmenté de plus de 7.200 milliards, à plus de 124.000 milliards. Le nombre d'hypothèses latentes est énorme et plus la production scientifique augmente en même temps que les scientifiques se spécialisent dans des domaines de plus en plus complexes, plus il y aura de connexions cachées. L'ordinateur ne génère pas d'hypothèse, mais en appui de la méthode de Swanson, il aide le chercheur à avancer rapidement dans le formidable volume de données auxquelles il peut accéder. C'est un des nouveaux outils qu'il a à sa disposition pour traiter différemment la littérature. A l'instar de l'ensemble des gènes qui forment le génome ou des protéines formant le protéome, la littérature scientifique constitue un ensemble de données liées entre elles (thèmes, auteurs, …) dont la quantité rend son exploitation difficile : certains parlent à juste titre de bibliome [Grivell, 2002]. Le traitement de grandes masses de données bibliographiques en appliquant la méthode transitive de Swanson peut mettre à profit cette quantité en révélant des données jusqu'alors latentes, données impossibles à obtenir par les moyens traditionnels d'exploitation des bases de données bibliographiques. Notre pratique nous a montré que l'approche de Swanson est un excellent stimulus pour la réflexion scientifique et produit des hypothèses robustes qui peuvent être intégré dans un processus de recherche et développement, n'attendant que la confrontation aux résultats expérimentaux.

Nous avons montré par ce travail que le DPM est une méthode efficace et rapide, combinant traitement de la littérature et expertise humaine pour générer des hypothèses. Cette méthode est modulaire et s'adapte à la nature du travail à effectuer : génération ou test d'hypothèse. Le DPM est fondé sur les travaux et la théorie de Don Swanson. Différentes équipes ont travaillé sur la théorie de Swanson, prenant ses travaux pour exemple. Les résultats obtenus conduisent toujours à formuler les mêmes hypothèses, malgré l'utilisation de techniques différentes. C'est là que réside la force du modèle de Don Swanson.

BIBLIOGRAPHIE

Afin de simplifier la lecture de la bibliographie de ce manuscrit, nous l'avons séparé en deux parties. La bibliographie générale, dont les références figurent en police normale [Swanson, 1986b], rapportée dans la section bibliographie. La bibliographie sur laquelle Don Swanson s'est appuyé pour ses travaux, dont les références figurent en italique *[Dyerberg*, *1982]* et sont présentées dans l'annexe 1.

Abbott, R. (2004), "Subjectivity as a concern for information science: a popperian perspective", Journal of Information Science. Vol. 30, n°2, p. 95-106.

Bawden, D. (2002), "The three worlds of health information", Journal of Information Science. Vol. 28, n°1, p. 51-62.

Blagosklonny, M.V., Pardee, A.B.(2002), "Unearthing the gems", Nature. Vol.416, n°6879, p. 373.

Blake, C., Pratt, W. (2002), "Automatically identifying candidate treatments from existing medical literature", in AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases. 2002. Palo Alto, California.

Blunt, R.J., George, A.J., Hurlow, R.A., Strachan, C.J., Stuart, J. (1980), "Hyperviscosity and thrombotic changes in idiopathic and secondary Raynaud's syndrome", British Journal of Haematology. Vol. 45, n°4, p. 651-658.

Brox, J.H., Killie, J.E., Gunnes, S., Nordoy, A. (1981), "The effects of cod liver oil and corn oil on platelets and vessel wall in man", Thrombosis and Haemostasis. Vol. 46, n°3, p. 604-611.

Bullinge, F. (2004), "Rôle et faiblesse de l'analyse dans la culture française de l'information", Regards sur l'IE. N° 5, p. 60-65.

Cadel, P., Boutin, E. (2005), "Les spécificités du Web : un obstacle à son exploitation ?", Communication, L'information numérique et les enjeux de la société de l'Information -ISD Tunis 14 au 16 avril 2005, 8 p.

Callon, M., Courtial, J.P., Laville, F. (1991), "Co-word analysis as a tool for describing the network of interaction between basic and technological research: the case of polymer chemistry", Scientometrics. Vol. 22, p. 155-205.

Catellin, S., (2003), "Sérendipité", Bulletin de la Société Française pour l'Histoire des Sciences de l'Homme. N°25, Automne-Hiver 2003, p. 27-32.

Cohen, A.M., Hersh, W.R. (2005), "A survey of current work in biomedical text mining", Briefings in Bioinformatics. Vol. 6, n°1, p. 57-71.

Coletti, M.H., Bleich, H.L. (2001), "Medical Subject Headings used to search the biomedical literature", Journal of the American Medical Informatics Association. Vol. 8, n°4, p. 317-323.

Demaine, J., Martin, J., De Bruiijn, B. (2003), "Haystacks and hypotheses", ASIST 2003 Annual Meeting - Humanizing Information Technology: From Ideas to Bits and Back, Westin Long Beach, California, 2003.

Demolombe, R., del Cerro, L.F. (1992), "Le raisonnement par abduction", Le Courrier du CNRS. N°79, p. 31.

Dewailly, E., Blanchet, C., Lemieux, S., Sauvé, L., Gingras, S., Ayotte, P ., Holub, B.J. (2001), "n-3 Fatty acids and cardiovascular disease risk factors among the Inuit of Nunavik", American Journal of Clinical Nutrition. Vol. 74, n°4, p. 948-954.

DiGiacomo, R.A., Kremer, J.M., Shah, D.M. (1989), "Fish-oil dietary supplementation in patients with Raynaud's phenomenon : a double-blind, controlled, prospective study", American Journal of Medicine. Vol. 86, n°2, p. 158-164.

DiMasi, J.A., Hansen, R.W, Grabowski, H.G. (2003), "The price of innovation: new estimates of drug development costs", Journal of Health Economics. Vol. 22, n°2, p. 151-185.

DiMasi, J.A., Grabowski, H.G., Vernon, J. (2004), "R&D costs and returns by therapeutic category", Drug Information Journal. Vol. 38, n°3, p. 211-223.

Dou, H., Hassanaly, P., Quoniam, L., La Tela, A. (1990), "Competitive technology assessment strategic patent clusters obtained with non-boolean logic. New application of the GET command", World Patent Information. Vol. 12, n°4, p. 222-229.

Driss, F., Vericel, E., Lagarde, M., Dechavanne, M., Darcet, P. (1984), "Inhibition of platelet aggregation and thromboxane synthesis after intake of small amount of icosapentaenoic acid", Thrombosis Research. Vol. 36, n°5, p. 389-396.

Edmunds, A., Morris, A. (2000), "The problem of information overload in business organisations: a review of the literature", International Journal of Information Management. Vol. 20, p. 17-28.

Fogarty, M., Bahls, C. (2002), "Information overload", The Scientist. Vol. 16, n°16, p. 16-18.

Fredholm, B.B., Abbracchio, M.P., Burnstock, G., Daly, J.W., Harden, T.K., Jacobson, K.A., Leff, P., Williams, M. (1994), "Nomenclature and classification of purinoreceptors", Pharmacological Reviews. Vol. 46, n°2, p. 143-156.

Gaarder, A., Jonsen, A., Laland, S., Hellem, A., Owren, P.A. (1961), "Adenosine diphosphate in red cells as a factor in the adhesiveness of human blood platelets", Nature. Vol. 192, p. 531-532.

Gachet, C., Cazenave, J.P., Ohlmann, P., Hilf, G., Wieland, T., Jacobs, K.H. (1992), "ADP receptor-induced activation of guanine-nucleotide-binding proteins in human platelet membranes", European Journal of Biochemistry. Vol. 207, n°1, p. 259-263.

Galland, S., Boulanger, N., Quoniam, L. (2004), "Soutenir l'innovation : l'impact des experts dans une démarche de veille stratégique", Actes du Colloque VSST 2004, Veille Stratégique Scientifique et Technologique, IRIT Toulouse, 25-29 octobre 2004. Textes des communications, Tome 1, p. 323-332.

Galland, S., Boulanger, N., Rostaing, H. (2005), "L'implication des experts dans un processus de prise de décision" Actes du 1er colloque européen d'Intelligence économique, ATELIS Poitiers, 27 & 28 janvier 2005, p. 61-67.

Goodman, A.B (1998), "Three independent lines of evidence suggest retinoids as causal to schizophrenia", Proceedings of the National Academy of Sciences of the United States of America. Vol. 95, n°13, p. 7240-7244.

Gordon, M.D., Lindsay, R.K. (1996), "Toward discovery support systems: a replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil", Journal of the American Society for Information Science. Vol. 47, n°2, p. 116-128.

Gordon, M.D., Dumais, S. (1998), "Using latent semantic indexing for literature based discovery", Journal of the American Society for Information Science. Vol. 49, n°8, p. 674-685.

Gordon, M., Lindsay, R.K, Fan, W. (2002), "Literature-based discovery on the World Wide Web", ACM Transactions on Internet Technology. Vol. 2, n°4, p. 261-275.

Grivell, L. (2002), "Mining the bibliome: searching for a needle in a haystack ? New computing tools are needed to effectively scan the growing amount of scientific literature for useful information", EMBO Reports. Vol. 3, n°3, p. 200-203.

Gulli, A., Signorini, A. (2005), "The indexable web is more than 11.5 billion pages", Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, ACM Press, New York (ISBN 1-59593-051-5), p. 902-903.

Hellem, A.J. (1960), "The adhesiveness of human blood platelets in vitro", Scandinavian Journal of Clinical and Laboratory Investigation. Vol. 12, n°12 (Suppl.), p. 1-117.

Hristovski, D., Stare, J., Peterlin, B., Dzeroski, S. (2001) " Supporting discovery in medicine by association rule mining in Medline and UMLS", Medinfo. Vol. 10, p. 1344-1348.

Hristovski, D., Peterlin, B., Mitchell, J.A., Humphrey, S.M. (2003), "Improving literature based discovery support by genetic knowledge integration", Studies in Health Technology and Informatics. Vol. 95, p. 68-73.

Hutton, R.A., Mikhailidis, D.P., Bernstein, R.M., Jeremy, J.Y., Hughes, G.R., Dandona P. (1984), "Assessment of platelet function in patients with Raynaud's syndrome", Journal of Clinical Pathology. Vol. 37, n°2, p. 182-187.

Josefson, D. (2001), "Second US institute investigates use of drug in asthma trial", British Medical Journal. Vol. 323, n°7308, p. 299.

Knoben, J.E., Phillips, S.J., Szczur, M.R. (2004), "The National Library of Medicine and drug information. Part 1: present ressources", Drug Information Journal. Vol. 38, n°1, p. 69-81.

Kostoff, R.N., DeMarco, R.A. (2001), "Extracting information from the literature by text mining", Analytical Chemistry. Vol. 73, n°13, p. 370A-378A.

Lachapelle, J.M. (2000), "Intérêt grandissant et limitations d'usage de la thalidomide en dermatologie", Louvain Médical. Vol. 119, n°9, p. S435-S439.

Lawrence, R.N. (2002), "Sir Richard Sykes contemplates the future of the pharma industry", Drug Discovery Today. Vol. 7, n°12, p. 645-648.

Lee, M.D., Weinblatt, M.E. (2001), "Rheumatoid arthritis", The Lancet. Vol. 358, n°9285, p. 903-911.

Lindsay, R.K., Gordon, M.D. (1999), "Literature-based discovery by lexical statistics", Journal of the American Society for Information Science. Vol. 50, n°7, p. 574-587.

Loh, S., Krug Wives, L., Palazzo M. de Oliveira, J. (2000), "Concept-based knowledge discovery in texts extracted from the web", SIGKDD Explorations. Vol. 2, n°1, p. 1-14.

Lyman, P., Varian, H.R., Swearingen, K., Charles, P., Good, N., Jordan, L.L., Pal, J. (2003), "How much information ? 2003", Study from the School of Information Management and Systems, University of California at Berkeley. (Page consultée le 22 septembre 2005). *How much information ? 2003*, [En ligne]. Adresse URL : http://www.sims.berkeley.edu/research/projects/how-much-info-2003/

Macfarlane, D.E., Mills, D.C.B. (1975), "The effect of ATP on platelets: evidence against the central role of released ADP in primary aggregation", Blood. Vol. 46, n°3, p. 309-320.

Mack, R., Hehenberger, M. (2002), "Text-based knowledge discovery: search and mining of life-sciences documents", Drug Discovery Today. Vol. 7, n°11 (Suppl.), p. S89-S98.

Marshall, E. (2001), "Human subjects. Volunteer's death prompts review", Science. Vol. 292, n°5525, p. 2226-2227.

Markellos, K., Perdikuri, K., Markello, P., Sirmakessis, S., Mayritsakis, G., Tsakalidis, A. (2002), "Knowledge discovery in patent databases", Proceedings of the 11th International Conference on Information and Knowledge Management (ISBN 1-58113-492-4), p. 672-674.

Mauskop, A., Altura, B.T., Cracco, R.Q., Altura, B.M. (1995), "Intravenous magnesium sulphate relieves migraine attacks in patients with low serum ionized magnesium levels: a pilot study", Clinical Science (London). Vol. 89, n°6, p. 633-636.

Mauskop, A., Altura, B.M. (1998), "Role of magnesium in the pathogenesis and treatment of migraines", Clinical Neuroscience. Vol. 5, n°1, p. 24-27.

McCarthy, M. (2001), "Healthy volunteer dies in US physiology study", The Lancet. Vol. 357, n°9274, p. 2114.

McLellan, F. (2001), "1966 and all that-when is a literature search done ?", The Lancet. Vol. 358, n°9282, p. 646.

Meek, I. (2004), "Drug discovery spending, a year in review", Drug Discovery World. Vol. 5, n°2, p. 9-14.

Mukherjea, S., Bamba, B., Kankar, P. (2005), "Information retrieval and knowledge discovery utilizing a biomedical patent semantic Web", IEEE Transactions on Knowledge Discovery and Data Mining. Vol. 17, n°8, p. 1099-1110.

Peikert, A., Wilimzig, C., Kohne-Volland, R. (1996) "Prophylaxis of migraine with oral magnesium: results from a prospective, multi-center, placebo-controlled and double-blind randomized study", Cephalalgia. Vol. 16, n°4, p. 257-263.

Perkins, E. (2001), "Johns Hopkins' tragedy : could librarians have prevented a death ?", Information Today. Vol. 18, n°8, p. 51.

Persidis, A., Deftereos, S., Persidis, A. (2004), "Systems literature analysis", Pharmacogenomics. Vol. 5, n°7, p. 943-947.

Pfaffenrath, V., Wessely, P., Meyer, C., Isler, H.R., Evers, S., Grotemeyer, K.H., Taneri, Z., Soyka, D., Gobel, H., Fischer, M. (1996), "Magnesium in the prophylaxis of migraine - a double-blind placebo-controlled study", Cephalalgia. Vol. 16, n°6, p. 436-440.

Pierret, J.D., Boutin E. (2004), "Découverte de connaissances dans les bases de données bibliographiques. Le travail de Don Swanson : de l'idée au modèle", ISDM. N°12, article n°109, 7p.

Pierret, J.D., Dolfi, F., Quoniam, L., Boutin E., Riccio, E.L. (2005), "Découverte de connaissances dans les bases de données bibliographiques. Modèles expérimentaux autour de la première hypothèse de Swanson", ISDM. N°20, article n°244, 12p.

Rikken, F., Vos , R. (1994), "Searching for adverse drug reactions at the margin of scientific fields", Scientometrics. Vol. 30, n°1, p. 187-199.

Rikken, F., Vos , R. (1995a), "Mapping the dynamics of adverse drug reactions in subsequent time periods using INDSCAL", Scientometrics. Vol. 33, n°3, p. 367-380.

Rikken, F., Vos , R. (1995b), "How adverse drug reactions can play a role in innovative drug research", Pharmacy World and Science. Vol. 17, n°6, p. 195-200.

Robillard, J. (2004), "Ontologies : antinomies, contradictions et autres difficultés épistémologiques du concept", La Revue STICEF. Vol. 11, 17 p. Sciences et Technologies de l´Information et de la Communication pour

l´Education et la Formation. (Page consultée le 22 septembre 2005). *Ontologies : antinomies, contradictions et autres difficultés épistémologiques du concept*, [En ligne]. Adresse URL : http://sticef.univ-lemans.fr/num/vol2004/robillard-05/sticef_2004_robillard_05.htm

Sandoli, D., Chiu, P.J., Chintala, M., Dionisotti, S., Ongini, E. (1994), "In vivo and ex vivo effects of adenosine A1 and A2 receptor agonists on platelet aggregation in the rabbit", European Journal of Pharmacolgy. Vol. 259, n°1, p. 43-49.

Schell, C.L., Rathe, R.J. (1992), "Meta-analysis – a tool for medical and scientific discovery", Bulletin of the Medical Library Association. Vol. 80, n°3, p. 219-222.

Shryock, J.C., Belardinelli, L. (1997), "Adenosine and adesosine receptors in the cardiovascular system: biochemistry, physiology, and pharmacology", American Journal of Cardiology. Vol. 79, n°12A, p. 2-10.

Silverman, W.A. (2002) "The schizophrenic career of a *monster drug*", Pediatrics. Vol. 110, n°2 Pt 1, p. 404-406.

Smaglik, P. (2001), "Asthma study death spurs inquiry", Nature. Vol. 411, n°6840, p. 873.

Smalheiser, N.R., Swanson D.R. (1994), "Assessing the gap in the biomedical literature: magnesium deficiency and neurologic disease", Neuroscience Research Communications. Vol. 15, n°1, p. 1-9.

Smalheiser, N.R., Swanson D.R. (1996a), "Indomethacin and Alzheimer's disease", Neurology. Vol. 46, n°2, p. 583.

Smalheiser, N.R., Swanson D.R. (1996b), "Linking estrogen to Alzheimer's disease: an informatics approach", Neurology. Vol. 47, n°3, p. 809-810.

Smalheiser, N.R., Swanson D.R. (1998a), "Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses", Computer Methods and Programs in Biomedicine. Vol. 57, n°3, p. 149-153.

Smalheiser, N.R., Swanson D.R. (1998b), "Calcium-independent phospholipase A2 and schizophrenia", Archives of General Psychiatry. Vol. 55, n°8, p. 752-753.

Smalheiser, N.R., Torvik, V.I., Weeber, M. and Swanson, D.R. (2002) "The Arrowsmith Project: New Tools to assist Biomedical Discovery and Collaboration", presented at the Beckman Institute, University of Illinois at Urbana-Champaign, October 1, 2002.

Spasser, M.A. (1997), "The enacted fate of undiscovery public knowledge", Journal of the American Society for Information Science. Vol. 48, n°8, p. 707-717.

Srinivasan, P. (2004), "Text mining: generating hypotheses from MEDLINE", Journal of the American Society for Information Science. Vol. 55, n°5, p. 396-413.

Stegmann, J., Grohmann, G. (2003), "Hypothesis generation guide by co-word clustering", Scientometrics. Vol. 56, n°1, p. 111-135.

Swanson, D.R. (1974), "Selective dissemination of biomedical information: a series of studies and model system", Library Quarterly. Vol. 44, n°3, p. 189-205.

Swanson, D.R. (1977), "Information retrieval as a trial-and-error process", Library Quarterly. Vol. 47, n°2, p. 128-148.

Swanson, D.R. (1979), "Libraries and the growth of knowledge", Library Quarterly. Vol. 49, n°1, p. 3-35.

Swanson, D.R. (1986a), "Fish oil, Raynaud's syndrome, and undiscovered public knowledge", Perspectives in Biology and Medicine. Vol. 30, n°1, p. 7-18.

Swanson, D.R. (1986b), "Undiscovered public knowledge", Library Quarterly. Vol. 56, n°2, p. 103-118.

Swanson, D.R. (1987), "Two medical literatures that are logically but not bibliographically connected", Journal of the American Society for Information Science. Vol. 38, n°4, p. 228-233.

Swanson, D.R. (1988), "Migraine and magnesium : eleven neglected connections", Perspectives in Biology and Medicine. Vol. 31, n°4, p. 526-557.

Swanson, D.R. (1989a), "Online search for logically-related noninteractive medical literatures : a systematic trial-and-error strategy", Journal of the American Society for Information Science. Vol. 40, n°5, p. 356-358.

Swanson, D.R. (1989b), "A second example of mutually isolated medical literatures related by implicit, unnoticed connections", Journal of the American Society for Information Science. Vol. 40, n°6, p. 432-435.

Swanson, D.R. (1990a), "Somatomedin C and arginin : implicit connections between mutually-isolated literatures", Perspectives in Biology and Medicine. Vol. 33, n°2, p. 157-186.

Swanson, D.R. (1990b), "Medical literature as a potential source of new knowledge", Bulletin of the Medical Library Association. Vol. 78, n°1, p. 29-37.

Swanson, D.R. (1993), "Intervening in the life cycles of scientific knowledge", Library Trends. Vol. 41, n°4, p. 606-631.

Swanson, D.R., Smalheiser, N.R. (1997), "An interactive system for finding complementary literatures: a stimulus to scientific discovery", Artificial Intelligence. Vol. 91, n°2 p. 183-203.

Swanson, D.R., Smalheiser, N.R. (1999), "Implicit text linkages between Medline records: using Arrowsmith as an aid to scientific discovery", Library Trends. Vol. 48, n°1, p. 48-59.

Swanson, D.R. (2001a), "ASIST Award of Merit acceptance speech : on fragmentation of knowledge, the connection explosion, and assembling other people's ideas", Bulletin of the American Society for Information Science and Technology. Vol. 27, n°3. ASIS. (Page consultée le 22 septembre 2005). *ASIST Award of Merit acceptance speech : on fragmentation of knowledge, the connection explosion, and assembling other people's ideas*, [En ligne]. Adresse URL : http://www.asis.org/Bulletin/Mar-01/swanson.html.

Swanson, D.R., Smalheiser, N.R., Bookstein A (2001b), "Information discovery from complementary literatures: categorizing viruses as potential weapons",

Journal of the American Society for Information Science and Technology. Vol. 52, n°10, p. 797-812.

Tabrizchi, R., Bedi, S. (2001), "Pharmacology of adenosine receptors in the vasculature", Pharmacology and Therapeutics. Vol. 91, n°2 p. 133-147.

Warne, P. (2003), "How drugs are developed: an introduction to pharmaceutical R&D", Scrip Reports BS1238. 108 p.

Weeber, M., Klein, H., Aronson, A.R., Mork, J.G., de Jong–van den Berg, L.T.W., Vos, R. (2000), "Text-based discovery in biomedicine: the architecture of the DAD-system", Proceedings of the AMIA Symposium. P. 903-907.

Weeber, M., Klein, H., de Jong-van den Berg, L.T.W, Vos, R. (2001), "Using concepts in literature-based discovery : simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries", Journal of the American Society for Information Science and Technology. Vol. 52, n°7, p. 548-557.

Weeber, M., Vos, R., Klein, H., de Jong–van den Berg, L.T.W., Aronson, A.R., Molema, G. (2003), "Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide", Journal of the American Medical Informatics Association. Vol. 10, n°3, p. 252-259.

Wren, J.D. (2004a), "The emerging in-silico scientist : how text-based bioinformatics is bridging biology and artificial intelligence", IEEE Engineering in Medicine and Biology Magazine. Vol. 23, n°2, p. 87-93.

Wren, J.D., Bekeredjian, R., Stewart, J.A., Shohet, R.V., Gerner, H.R. (2004b), "Knowledge discovery by automated identification and ranking of implicit relationships", Bioinformatics. Vol. 20, n°3, p. 389-398.

ANNEXE 1

Bibliographie supportant

la première découverte de Swanson

Afin de simplifier la lecture de la bibliographie de ce manuscrit, nous l'avons séparé en deux parties. La bibliographie générale, dont les références figurent en police normale [Swanson, 1986b], rapportée dans la section bibliographie. La bibliographie sur laquelle Don Swanson s'est appuyé pour ses travaux, dont les références figurent en italique *[Dyerberg*, *1982]* et sont présentées dans l'annexe 1.

A1.1 Bibliographie sur la maladie de Raynaud – 34 articles

Belch, J.J., Newman, P., Drury, J.K., McKenzie, F., Capell, H., Leiberman, P, Forbes, C.D., Prentice, C.R. (1983), "Intermittent epoprostenol (prostacyclin) infusion in patients with Raynaud's syndrome. A double-blind controlled trial", The Lancet. Vol. 1, n°8320, p. 313-315.

Belch, J.J., McLaren, M., Anderson, J., Lowe, G.D., Sturrock, R.D., Capell, H.A., Forbes, C.D. (1985a), "Increased prostacyclin metabolites and decreased red cell deformability in patients with systemic sclerosis and Raynaud's syndrome", Prostaglandins, Leukotrienes, and Medicine. Vol. 17, n°1, p. 1-9.

Blunt, R.J., George, A.J., Hurlow, R.A., Strachan, C.J., Stuart, J. (1980), "Hyperviscosity and thrombotic changes in idiopathic and secondary Raynaud's syndrome", British Journal of Haematology. Vol. 45, n°4, p. 651-658.

Bounameaux, H.M., Hellemans, H., Verhaeghe, R., Dequeker, J. (1984), "Ketanserin (5 HT2-antagonist) in secondary Raynaud's phenomenon", Journal of Cardiovascular Pharmacology. Vol. 6, n°5, p. 975-976.

Dowd, P.M., Kovacs, I.B., Bland, C.J., Kirby, J.D. (1981), "Effect of prostaglandins I2 and E1 on red cell deformability in patients with Raynaud's

phenomenon and systemic sclerosis", British Medical Journal (Clinical research ed.). Vol. 283, n°6287, p. 350.

Dowd, P.M., Martin, M.F., Cooke, E.D., Bowcock, S.A., Jones, R., Dieppe, P.A., Kirby, J.D. (1982), "Treatment of Raynaud's phenomenon by intravenous infusion of prostacyclin (PGI2)", British Journal of Dermatololgy. Vol. 106, n°1, p. 81-89.

Goyle, K.B., Dormandy, J.A. (1976), "Abnormal blood viscosity in Raynaud's phenomenon", The Lancet. Vol. 1, n°7973, p. 1317-1318.

Horrobin, D.F., Jenkins, K., Manku, M.S. (1983), "Raynaud's phenomenon, histamine, and prostaglandins", The Lancet. Vol. 2, n°8352, p. 747-748.

Hutton, R.A., Mikhailidis, D.P., Bernstein, R.M., Jeremy, J.Y., Hughes, G.R., Dandona, P. (1984), "Assessment of platelet function in patients with Raynaud's syndrome", Journal of Clinical Pathology. Vol. 37, n°2, p. 182-187.

Kahan, A., Weber, S., Amor, B., Menkes, C.J., Saporta, L., Hodara, M., Guerin, F., Degeorges, M. (1983), "Calcium entry blocking agents in digital vasospasm (Raynaud's phenomenon)", European Heart Journal. Vol. 4 Suppl C, p. 123-129.

Kallenberg, C.G., Vellenga, E., Wouda, A.A., The, T.H. (1982), "Platelet activation, fibrinolytic activity and circulating immune complexes in Raynaud's phenomenon", Journal of Rheumatology. Vol. 9, n°6, p. 878-884.

Keller, J., Kaltenecker, A., Schricker, K.T., Neidhardt, B., Hornstein, O.P. (1985), "Inhibition of platelet aggregation by a new stable prostacyclin introduced in therapy of patients with progressive scleroderma", Archives of Dermatological Research. Vol. 277, n°4, p. 323-325.

Kovacs, I.B., O'Grady, J. (1984), "Prostacyclin increases filterability of normal and rigidified human red blood cells in vitro", Agents and Actions. Vol. 14, n°2, p. 306-310.

Kyle, V., Parr, G., Salisbury, R., Thomas, P.P., Hazleman, B. (1985), "Prostaglandin E1 vasospastic disease and thermography", Annals of the Rheumatic Diseases. Vol. 44, n°2, p. 73-78.

Larcan, A., Schmidt, C., Stoltz, J.F., Voisin, P. (1984), "Hémorhéologie des syndromes de Raynaud", Journal des Maladies Vasculaires. Vol. 9, n°1, p. 1-6.

Longstaff, J., Gush, R., Williams, E.H., Jayson, M.I. (1985), "Effects of ketanserin on peripheral blood flow, haemorheology, and platelet function in patients with Raynaud's phenomenon", Journal of Cardiovascular Pharmacology. Vol. 7 Suppl 7, p. S99-101.

Lucas, G.S., Simms, M.H., Caldwell, N.M., Alexander, S.J., Stuart, J. (1984), "Haemorrheological effects of prostaglandin E1 infusion in Raynaud's syndrome", Journal of Clinical Pathology. Vol. 37, n°8, p. 870-873.

Malamet, R., Wise, R.A., Ettinger, W.H., Wigley, F.M. (1985), "Nifedipine in the treatment of Raynaud's phenomenon. Evidence for inhibition of platelet activation", American Journal of Medicine. Vol. 78, n°4, p. 602-608.

Martin, M.F., Dowd, P.M., Ring, E.F., Cooke, E.D., Dieppe, P.A., Kirby, J.D. (1981), "Prostaglandin E1 infusions for vascular insufficiency in progressive systemic sclerosis", Annals of the Rheumatic Diseases. Vol. 40, n°4, p. 350-354.

McGrath, M.A., Peek, R., Penny, R. (1977), "Blood hyperviscosity with reduced skin blood flow in scleroderma", Annals of the Rheumatic Diseases. Vol. 36, n°6, p. 569-574.

McGrath, M.A., Peek, R., Penny, R. (1978), "Raynaud's disease: reduced hand blood flows with normal blood viscosity", Australian and New Zealand Journal of Medicine. Vol. 8, n°2, p. 126-131.

Pardy, B.J., Hoare, M.C., Eastcott, H.H., Miles, C.C., Needham, T.N., Harbourne, T., Ellis, B.W. (1982), "Prostaglandin E1 in severe Raynaud's phenomenon", Surgery. Vol. 92, n°6, p. 953-965.

Pola, P., Savi, L., Dal Lago, A., Flore, R., Shami, J. (1980), "Invariability of blood viscosity after cold testing in patients suffering for Raynaud's disease.", Journal of Cardiovascular Surgery. Vol. 21, n°2, p. 211-214.

Pringle, R., Walder, D.N. Weaver, J.P. (1965), "Blood viscosity and Raynaud's disease", The Lancet. Vol. 14, p. 1086-1088.

Roald, O.K., Seem, E. (1984), "Treatment of Raynaud's phenomenon with ketanserin in patients with connective tissue disorders.", British Medical Journal (Clinical research ed.). Vol. 289, n°6445, p. 577-579.

Rodeheffer, R.J., Rommer, J.A., Wigley, F., Smith, C.R. (1983), "Controlled double-blind trial of nifedipine in the treatment of Raynaud's phenomenon", N Engl J Med. Vol. 308, n°15, p. 880-883.

Sandhagen, B., Wegener, T. (1985), "Blood viscosity and finger systolic pressure in primary and traumatic vasospastic disease", Upsala Journal of Medical Sciences. Vol. 90, n°1, p. 55-59.

Sauza, J., Kraus, A., Gonzalez-Amaro, R., Alarcon-Segovia, D. (1984), "Effect of the calcium channel blocker nifedipine on Raynaud's phenomenon. A controlled double blind trial", Journal of Rheumatology. Vol. 11, n°3, p. 362-364.

Seibold, J.R., Jageneau, A.H. (1984), "Treatment of Raynaud's phenomenon with ketanserin, a selective antagonist of the serotonin2 (5-HT2) receptor", Arthritis and Rheumatism. Vol. 27, n°2, p. 139-146.

Smith, C.D., McKendry, R.J. (1982), "Controlled trial of nifedipine in the treatment of Raynaud's phenomenon.", The Lancet. Vol. 2, n°8311, p. 1299-1301.

Stranden, E., Roald, O.K., Krohg, K. (1982), "Treatment of Raynaud's phenomenon with the 5-HT2-receptor antagonist ketanserin", British Medical Journal (Clinical research ed.). Vol. 285, n°6348, p. 1069-1071.

Tietjen, G.W., Chien, S., Leroy, E.C., Gavras, I., Gavras, H., Gump, F.E. (1975), "Blood viscosity, plasma proteins, and Raynaud syndrome", Archives of Surgery. Vol. 110, n°11, p. 1343-1346.

Walker, R.T., Matrai, A., Bogar, L., Dormandy, J.A. (1985), "Serotonin and the flow properties of blood.", Journal of Cardiovascular Pharmacology. Vol. 7 Suppl 7, p. S35-S37.

Zahavi J., Hamilton W.A., O'Reilly M.J., Leyton J., Cotton L.T., Kakkar V.V. (1980), "Plasma exchange and platelet function in Raynaud's phenomenon", Thrombosis Research. Vol. 19, n°1-2, p. 85-93.

A1.2 Bibliographie sur l'huile de poisson – 25 articles

Brox, J.H., Killie, J.E., Gunnes, S., Nordoy, A. (1981), "The effect of cod liver oil and corn oil on platelets and vessel wall in man", Thrombosis and Haemostasis. Vol. 46, n°3, p. 604-611.

Brox, J.H., Killie, J.E., Osterud, B., Holme, S., Nordoy, A. (1983), "Effects of cod liver oil on platelets and coagulation in familial hypercholesterolemia (type IIa)", Acta Medica Scandinavica. Vol. 213, n°2, p. 137-144.

Cartwright, I.J., Pockley, A.G., Galloway, J.H., Greaves, M., Preston, F.E. (1985), "The effects of dietary omega-3 polyunsaturated fatty acids on erythrocyte membrane phospholipids, erythrocyte deformability and blood viscosity in healthy volunteers", Atherosclerosis. Vol. 55, n°3, p. 267-281.

Driss, F., Vericel, E., Lagarde, M., Dechavanne, M., Darcet, P. (1984), "Inhibition of platelet aggregation and thromboxane synthesis after intake of small amount of icosapentaenoic acid", Thrombosis Research. Vol. 36, n°5, p. 389-396.

Dyerberg, J., Jorgensen, K.A. (1982), "Marine oils and thrombogenesis.", Progress in Lipid Research. Vol. 21, n°4, p.255-269.

Fehily, A.M., Burr, M.L., Phillips, K.M., Deadman, N.M. (1983), "The effect of fatty fish on plasma lipid and lipoprotein concentrations", American Journal of Clinical Nutrition. Vol. 38, n°3, p. 349-351.

Fischer, S., Weber, P.C. (1984), "Prostaglandin I3 is formed in vivo in man after dietary eicosapentaenoic acid", Nature. Vol. 307, n°5947, p. 165-168.

Goodnight, S.H. Jr, Harris, W.S., Connor, W.E. (1981), "The effects of dietary omega 3 fatty acids on platelet composition and function in man: a prospective, controlled study", Blood. Vol. 58, n°5, p. 880-885.

Harris, W.S., Connor, W.E., McMurry, M.P. (1983), "The comparative reductions of the plasma lipids and lipoproteins by dietary polyunsaturated fats: salmon oil versus vegetable oils", Metabolism: Clinical and Experimental. Vol. 32, n°2, p. 179-184.

Hashimoto, Y., Naito, C., Kawamura, M., Oka, H. (1984), "Effects of the ratio of exogenous eicosapentaenoic acid to arachidonic acid on platelet aggregation and serotonin release", Thrombosis Research. Vol. 34, n°5, p. 439-446.

Kristensen, S.D., Arnfred, T., Dyerberg, J. (1984), "Eicosapentaenoic acid potentiates the production of prostacyclin-like material in the arachidonic acid perfused human umbilical vein", Thrombosis Research. Vol. 36, n°4, p. 305-314.

Lockette, W.E., Webb, R.C., Culp, B.R., Pitt, B. (1982), "Vascular reactivity and high dietary eicosapentaenoic acid", Prostaglandins. Vol. 24, n°5, p. 631-639.

Morita, I., Takahashi, R., Saito, Y., Murota, S. (1983), "Effects of eicosapentaenoic acid on arachidonic acid metabolism in cultured vascular cells and platelets: species difference", Thrombosis Research. Vol. 31, n°2, p. 211-217.

Mortensen, J.Z., Schmidt, E.B., Nielsen, A.H., Dyerberg, J. (1983), "The effect of N-6 and N-3 polyunsaturated fatty acids on hemostasis, blood lipids and blood pressure.", Thrombosis and Haemostasis. Vol. 50, n°2, p. 543-546.

Nagakawa, Y., Orimo, H., Harasawa, M., Morita, I., Yashiro, K., Murota, S. (1983), "Effect of eicosapentaenoic acid on the platelet aggregation and composition of fatty acid in man. A double blind study", Atherosclerosis. Vol. 47, n°1, p. 71-75.

Phillipson, B.E., Rothrock, D.W., Connor, W.E., Harris, W.S., Illingworth, D.R. (1985), "Reduction of plasma lipids, lipoproteins, and apoproteins by dietary fish oils in patients with hypertriglyceridemia", New England Journal of Medicine. Vol. 312, n°19, p. 1210-1216.

Sanders, T.A., Hochland, M.C. (1983), "A comparison of the influence on plasma lipids and platelet function of supplements of omega 3 and omega 6 polyunsaturated fatty acids", British Journal of Nutrition. Vol. 50, n°3, p. 521-529.

Saynor, R., Verel, D. (1983), "Eskimos and their diets", The Lancet. Vol. 1, n°8337, p. 1335.

Saynor, R., Verel, D., Gillott, T. (1984), "The long-term effect of dietary supplementation with fish lipid concentrate on serum lipids, bleeding time, platelets and angina", Atherosclerosis. Vol. 50, n°1, p. 3-10.

Simons, L.A., Hickie, J.B., Balasubramaniam, S. (1985), "On the effects of dietary n-3 fatty acids (Maxepa) on plasma lipids and lipoproteins in patients with hyperlipidaemia", Atherosclerosis. Vol. 54, n°1, p. 75-88.

Singer, P., Wirth, M., Voigt, S., Zimontkowski, S., Godicke, W., Heine, H. (1984), "Clinical studies on lipid and blood pressure lowering effect of eicosapentaenoic acid-rich diet", Biomed Biochim Acta. Vol. 43, n°8-9, p. S421-S425.

Spector, A.A., Kaduce, T.L., Figard, P.H., Norton, K.C., Hoak, J.C., Czervionke, R.L. (1983), "Eicosapentaenoic acid and prostacyclin production by cultured human endothelial cells.", Journal of Lipid Research. Vol. 24, n°12, p. 1595-1604.

Terano, T., Hirai, A., Hamazaki, T., Kobayashi, S., Fujita, T., Tamura, Y., Kumagai, A. (1983), "Effect of oral administration of highly purified eicosapentaenoic acid on platelet function, blood viscosity and red cell deformability in healthy human subjects", Atherosclerosis. Vol. 46, n°3, p. 321-331.

Von Schacky, C., Fischer, S., Weber, P.C. (1985), "Long-term effects of dietary marine omega-3 fatty acids upon plasma and cellular lipids, platelet function, and eicosanoid formation in humans", Journal of Clinical Investigation. Vol. 76, n°4, p. 1626-1631.

Woodcock, B.E., Smith, E., Lambert, W.H., Jones, W.M., Galloway, J.H., Greaves, M., Preston, F.E. (1984), "Beneficial effect of fish oil on blood viscosity in peripheral vascular disease", British Medical Journal (Clinical research ed.). Vol. 288, n°6417, p. 592-594.

A1.3 Bibliographie complémentaire (articles cités, couplage …)

Belch, J.J., Shaw, B., O'Dowd, A., Saniabadi, A., Leiberman, P., Sturrock, R.D., Forbes, C.D. (1985b), "Evening primrose oil (Efamol) in the treatment of Raynaud's phenomenon: a double blind study", Thrombosis and Haemostasis. Vol. 54, n°2, p. 490-494.

Dormandy, J.A., Hoare, E., Colley, J., Arrowsmith, D.E., Dormandy, T.L. (1973), "Clinical, haemodynamic, rheological, and biochemical findings in 126 patients with intermittent claudication", British Medical Journal. Vol. 4, n°5892, p. 576-581.

Horrobin, D.F. (1984), "Essential fatty acid metabolism in diseases of connective tissue with special reference to scleroderma and to Sjogren's syndrome", Medical Hypotheses. Vol. 14, n°3, p. 233-247.

Hung, S.C., Ghali, N.I., Venton, D.L., Le Breton, G.C. (1982), " Prostaglandin F2 alpha antagonizes thromboxane A2-induced human platelet aggregation", Prostaglandins. Vol. 24, n°2, p. 195-206.

Moncada, S. (1983), "Biology and therapeutic potential of prostacyclin", Stroke. Vol. 14, n°2, p.157-168.

Moncada, S., Vane, J.R. (1984), "Prostacyclin and its clinical applications", Annals of Clinical Research. Vol. 16, n°5-6, p. 241-252.

Ozanne, P., Boudart, D., Mainard, F., Lefebvre, J., Grolleau, J.Y. (1984), "Hyperviscosité sanguine et plasmatique au cours des hyperlipidémies primitives", Revue de Médecine Interne. Vol. 5, n°1, p. 29-33.

Preston, F.E., Greaves, M. (1985), "Platelet suppressive therapy in clinical medicine", British Journal of Haematology. Vol. 60, n°4, p.589-597.

Seplowitz, A.H., Chien, S., Smith, F.R. (1981), "Effects of lipoproteins on plasma viscosity", Atherosclerosis. Vol. 38, n°1-2, p. 89-95.

Weksler B.B. (1984), "Prostaglandins and vascular function", Circulation. Vol. 70, n°5 Pt 2, p. III63-71.

ANNEXE 2

Exemple de citation Medline

PMID- 2403828
OWN - NLM
STAT- MEDLINE
DA - 19900221
DCOM- 19900221
LR - 20041117
PUBM- Print
IS - 0025-7338
VI - 78
IP - 1
DP - 1990 Jan
TI - Medical literature as a potential source of new knowledge.
PG - 29-37
AB - Specialized biomedical literatures have been found that are implicitly linked by arguments that they respectively contain, but which nonetheless do not cite or refer to each other. The combined arguments lead to new inferences and conclusions that cannot be drawn from the separate literatures. One such analysis identified one set of articles showing that dietary fish oils lead to certain blood and vascular changes, and a second set containing evidence that similar changes might benefit patients with Raynaud's syndrome. Yet these two literatures had no articles in common and had never before been cited together; neither literature mentioned the other or suggested that dietary fish oil might benefit Raynaud patients. Two years after publication of that analysis, the first clinical trial demonstrating such a beneficial effect was reported independently by others. A second example of literature synthesis, based on eleven indirect connections, led to an inference that magnesium deficiency might be a causal factor in migraine headache. A third example calls attention to implicit connections between arginine intake and blood levels of somatomedins, a potentially fruitful but neglected area of research with implications for the decline with age of thymic function and protein synthesis. A model and an online search strategy to aid in identifying other logically related noninteractive literatures is described. Such structures are probably not rare and may provide the foundation for a literature-based approach to scientific discovery.
AD - Graduate Library School, Center for Information Studies, University of Chicago, IL 60637.
FAU - Swanson, D R
AU - Swanson DR
LA - eng
PT - Journal Article
PL - UNITED STATES
TA - Bull Med Libr Assoc
JID - 0421037
RN - 0 (Fish Oils)

```
RN  - 67763-96-6 (Insulin-Like Growth Factor I)
RN  - 74-79-3 (Arginine)
SB  - IM
MH  - Abstracting and Indexing/methods
MH  - Arginine/pharmacology
MH  - Artificial Intelligence
MH  - Fish Oils/therapeutic use
MH  - Forecasting
MH  - Humans
MH  - Insulin-Like Growth Factor I/metabolism
MH  - MEDLARS
MH  - Magnesium Deficiency/complications
MH  - Migraine/etiology
MH  - *Online Systems/organization & administration
MH  - *Periodicals
MH  - Raynaud Disease/diet therapy
MH  - Research Support, U.S. Gov't, Non-P.H.S.
MH  - United States
EDAT- 1990/01/01
MHDA- 1990/01/01 00:01
PST - ppublish
SO  - Bull Med Libr Assoc 1990 Jan;78(1):29-37.
```

(Citation reproduite avec la permission de la NLM et de la Medical Library Association, Chicago)

PubMed propose un lien gratuit vers le texte intégral de cet article.

ANNEXE 3

Dictionnaires DPM selon le MeSH 2005

A3.1 Tree Drugs (tree drugs.txt) includes :

D01 to D27 (Note : exclude proteins).

Details

Inorganic Chemicals [D01]
Organic Chemicals [D02]
Heterocyclic Compounds [D03]
Polycyclic Compounds [D04]
Macromolecular Substances [D05] (excluding D05.500)
Hormones, Hormone Substitutes, and Hormone Antagonists [D06]
Reproductive Control Agents [D07]
Enzymes and Coenzymes [D08] (excluding D08.244 - D08.622 - D08.811)
Carbohydrates [D09]
Lipids [D10]
Growth Substances, Pigments, and Vitamins [D11] (excluding D11.303.330 – D11.303.580)
Amino Acids, Peptides, and Proteins [D12] (excluding D12.776)
Nucleic Acids, Nucleotides, and Nucleosides [D13] (excluding (D13.150 – D13.400 – D13.444)
Neurotransmitters and Neurotransmitter Agents [D14]
Central Nervous System Agents [D15]
Peripheral Nervous System Agents [D16]
Anti-Inflammatory Agents, Antirheumatic Agents, and Inflammation Mediators [D17]
Cardiovascular Agents [D18]
Hematologic, Gastrointestinal, and Renal Agents [D19]
Complex Mixtures [D20]
Anti-Allergic and Respiratory System Agents [D21]
Antineoplastic and Immunosuppressive Agents [D22]
Dermatologic Agents [D23]
Immunologic and Biological Factors [D24] (excluding D24.611.125 - D24.611.171 - D24.611.216)
Biomedical and Dental Materials [D25]
Pharmaceutical Preparations [D26]
Chemical Actions and Uses [D27]

A3.2 Tree Proteins/Targets (tree prot.txt) includes (Note : include DNA/RNA) :

D05.500
D08.244 - D08.622 - D08.811
D11.303.330 - D11.303.580
D12.776
D13.150 - D13.400 - D13.444
D24.611.125 - D24.611.171 - D24.611.216

Details

D05.500 – Multiprotein Complexes
D08.244 - Cytochromes
D08.622 - Enzyme Precursors
D08.811 - Enzymes
D11.303.330 - Cyclins
D11.303.580 - Maturation-Promoting Factor
D12.776 - Proteins
D13.150 - Antisense Elements (Genetics)
D13.400 - Nucleic Acid Precursors
D13.444 - Nucleic Acids
D24.611.125 - Antibodies
D24.611.171 - Antigen-Antibody Complex
D24.611.216 - Antigens

A3.3 Tree Physiology (tree physio.txt) includes :

F02 - G04 to G11 – G13 – G14

Details

F02 - Psychological Phenomena and Processes
G04 - Biological Phenomena, Cell Phenomena, and Immunity
G05 - Genetic Processes
G06 - Biochemical Phenomena, Metabolism, and Nutrition
G07 - Physiological Processes
G08 - Reproductive and Urinary Physiology
G09 - Circulatory and Respiratory Physiology
G10 - Digestive, Oral, and Skin Physiology
G11 - Musculoskeletal, Neural, and Ocular Physiology
G13 - Genetic Phenomena
G14 - Genetic Structures

A3.4 Tree Diseases (tree disea.txt) includes :

C01 to C23 – F03

Details

C01 - Bacterial Infections and Mycoses
C02 - Virus Diseases
C03 - Parasitic Diseases
C04 - Neoplasms
C05 - Musculoskeletal Diseases
C06 - Digestive System Diseases
C07 - Stomatognathic Diseases
C08 - Respiratory Tract Diseases
C09 - Otorhinolaryngologic Diseases
C10 - Nervous System Diseases
C11 - Eye Diseases
C12 - Urologic and Male Genital Diseases
C13 - Female Genital Diseases and Pregnancy Complications
C14 - Cardiovascular Diseases
C15 - Hemic and Lymphatic Diseases
C16 - Neonatal Diseases and Abnormalities
C17 - Skin and Connective Tissue Diseases
C18 - Nutritional and Metabolic Diseases
C19 - Endocrine Diseases
C20 - Immunologic Diseases
C21 - Disorders of Environmental Origin
C22 - Animal Diseases
C23 - Pathological Conditions, Signs and Symptoms
F03 - Mental Disorders

A3.5 Tree Anatomy (tree anat.txt) includes :

A01 to A17

Details

A01 - Body Regions
A02 - Musculoskeletal System
A03 - Digestive System
A04 - Respiratory System
A05 - Urogenital System
A06 - Endocrine System
A07 - Cardiovascular System
A08 - Nervous System
A09 - Sense Organs
A10 - Tissues
A11 - Cells
A12 - Fluids and Secretions
A13 - Animal Structures
A14 - Stomatognathic System
A15 - Hemic and Immune Systems
A16 - Embryonic Structures
A17 - Integumentary System

A3.6 Tree Dietary Factors (tree diet.txt) includes :

D01.248 – D01.268 – D01.496 – D09.301 – D10 – D11.430 – D11.786

Details

Electrolytes [D01.248]
Elements [D01.268]
Isotopes [D01.496]
Dietary Carbohydrates [D09.301]
Lipids [D10]
Micronutrients [D11.430]
Vitamins [D11.786]

ANNEXE 4

Liste des concepts B – physiologie

Premier et deuxième DPM

maladie de Raynaud/huile de poisson

Voir paragraphes *2.3.3* et *2.4.1 Extraction des concepts B*

Descripteur MeSH	Freq
Skin Temperature	92
Regional Blood Flow	86
Blood Pressure	70
Blood Viscosity	41
Vasoconstriction	34
Blood Flow Velocity	32
Biofeedback (Psychology)	30
Body Temperature	30
Blood Circulation	26
Microcirculation	26
Hemodynamic Processes	21
Vasodilation	19
Pulse	16
Body Temperature Regulation	14
Necrosis	13
Bone Resorption	11
Pain	11
Vascular Resistance	11
Heart Rate	10
Pregnancy	10
Platelet Aggregation	8
Fibrinolysis	7
Stress, Psychological	7
Blood Coagulation	6
Hematocrit	6
Reflex	6
Centromere	5
Sweating	5
Collateral Circulation	4
Differential Threshold	4
Gastrointestinal Motility	4
Laterality	4
Movement	4
Muscle Relaxation	4
Osteolysis	4
Pulmonary Diffusing Capacity	4
Systole	4
Cardiac Output	3
Chromosomes	3
Conditioning, Classical	3
Conditioning, Operant	3
Electrophysiology	3
Erythrocyte Aggregation	3
Erythrocyte Deformability	3
Evoked Potentials	3
Galvanic Skin Response	3
Muscle Contraction	3
Peristalsis	3
Platelet Adhesiveness	3
Posture	3
Sensation	3
Sensory Thresholds	3
Temperature Sense	3
Wound Healing	3
Aging	2
Antigen-Antibody Reactions	2
Arousal	2
Capillary Permeability	2
Cerebrovascular Circulation	2
Exertion	2
Hemostasis	2
Menstruation	2
Muscle Tonus	2
Nerve Regeneration	2
Neural Conduction	2
Oxygen Consumption	2

Prothrombin Time	2
Pulmonary Circulation	2
Remission, Spontaneous	2
Skin Physiology	2
Sleep	2
Stroke Volume	2
Touch	2
Action Potentials	1
Adaptation, Biological	1
Adaptation, Physiological	1
Antibody Specificity	1
Body Image	1
Body Weight	1
Cell Movement	1
Chromosome Aberrations	1
Chromosomes, Human, 1-3	1
Chromosomes, Human, 6-12 and X	1
Deglutition	1
Erythrocyte Count	1
Evoked Potentials, Somatosensory	1
Gastric Emptying	1
Generalization (Psychology)	1
Generalization, Response	1
Histamine Release	1

Intestinal Absorption	1
Lactation	1
Lung Compliance	1
Memory	1
Menopause	1
Metaphase	1
Motor Skills	1
Myocardial Contraction	1
Oxidation-Reduction	1
Platelet Count	1
Pregnancy Trimester, Second	1
Pulmonary Gas Exchange	1
Reaction Time	1
Reflex, Abnormal	1
Reflex, Stretch	1
Respiration	1
Self Stimulation	1
Space Perception	1
Swimming	1
Synaptic Transmission	1
Taste	1
Twins, Monozygotic	1
Valsalva Maneuver	1
Visual Perception	1
Water-Electrolyte Balance	1

ANNEXE 5

Liste des concepts A – diet factors

Premier DPM

maladie de Raynaud/huile de poisson

Voir paragraphe *2.3.5 Extraction des concepts A*

Descripteur MeSH	Freq
Arachidonic Acids	516
Epoprostenol	467
Prostaglandins	464
Calcium	441
Lipids	321
Cholesterol	269
Phospholipids	224
Prostaglandins E	220
Arachidonic Acid	214
Thromboxane B2	194
Thromboxanes	190
Oxygen	172
Thromboxane A2	166
Triglycerides	152
Radioisotopes	131
Xenon	117
Platelet Activating Factor	115
Dietary Fats	108
Lipoproteins	101
Fatty Acids	92
Vitamin K	88
Fatty Acids, Nonesterified	82
Fatty Acids, Unsaturated	82
Magnesium	76
Sodium	74
Iodine Radioisotopes	70
Carbon Radioisotopes	67
Iodine Isotopes	67
Platelet Factor 3	64
Alprostadil	61
Prostaglandin Endoperoxides, Synthetic	59
Phosphatidyl-ethanolamines	57
Prostaglandins F	56
Nicotinic Acids	55
Lysophosphatidylcholines	51
Prostaglandins D	51
Carbon Isotopes	50
Phosphatidylcholines	49
Prostaglandins, Synthetic	49
6-Ketoprostaglandin F1 alpha	48
Potassium	47
Prostaglandin Endoperoxides	44
Prostaglandins H	41
Lipopolysaccharides	38
Vitamin E	37
Krypton	36
Linoleic Acids	36
Lipoproteins, LDL	36
Tritium	36
Chromium Radioisotopes	35
Eicosapentaenoic Acid	35
Propionates	33
Serum Albumin, Radio-Iodinated	33
Rubidium	32
Indium	31
Peroxides	31
15-Hydroxy-11 alpha,9 alpha-(epoxymethano)prosta-5,13-dienoic Acid	30
Dinoprostone	29

Lipoproteins, HDL	29
Phosphatidylinositols	28
Ascorbic Acid	27
Membrane Lipids	27
Sulfates	26
Prostaglandins E, Synthetic	25
Cations, Divalent	22
Oils	22
Prostaglandin D2	22
Chromium Isotopes	21
Hydrogen Peroxide	21
Lipoproteins, VLDL	21
Carbon	20
Dinoprost	20
Cyanides	19
Bicarbonates	18
Chlorides	18
Cobalt	18
Hydrogen	18
Iron	18
Phosphatidic Acids	18
Zinc	18
Acetic Acids	17
Butyrates	17
Iloprost	17
Phosphatidylserines	17
Copper	16
Electrolytes	16
Fatty Acids, Essential	16
Ions	16
Prostanoic Acids	16
Eicosanoic Acids	15
Phosphorus Isotopes	15
Fish Oils	14
Linolenic Acids	14
Lithium	14
Manganese	14
Nitrogen	14
Technetium	14
Iodine	13
Prostaglandin H2	13
Cholesterol, Dietary	12
Glycerides	12
Lipid Peroxides	12
Lipoproteins, HDL Cholesterol	12
Nitroprusside	12
Oleic Acids	12
Sphingomyelins	12
Strontium	12
5,8,11,14-Eicosatetraynoic Acid	11
Barium	11
Butter	11
Gold	11

Phosphates	11
Prostaglandins F, Synthetic	11
Fats, Unsaturated	10
Vitamins	10
8,11,14-Eicosatrienoic Acid	9
Diglycerides	9
Lysophospholipids	9
Phosphorus	9
Pyridoxal Phosphate	9
Stearic Acids	9
Sulfides	9
Vitamin K 1	9
Carbonates	8
Cod Liver Oil	8
Dietary Carbohydrates	8
Mercury	8
Nickel	8
Nitrates	8
Sulfoglycosphingolipids	8
Vitamin B 12	8
Vitamin B Complex	8
12-Hydroxy-5,8,10,14-eicosatetraenoic Acid	7
Aminobenzoic Acids	7
Cadmium	7
Charcoal	7
Ferricyanides	7
Fluorides	7
Gold Colloid, Radioactive	7
Phosphorus Radioisotopes	7
Propionic Acids	7
Prostaglandins G	7
Pyridoxine	7
Sodium Dodecyl Sulfate	7
Vitamin A	7
Calcium Radioisotopes	6
Chlorine	6
Chylomicrons	6
Cobalt Isotopes	6
Disulfides	6
Fluorine	6
Lanthanum	6
Palmitic Acids	6
Potassium Isotopes	6
Strontium Isotopes	6
Xenon Radioisotopes	6
Anions	5
Bromine	5
Cobalt Radioisotopes	5
Docosahexaenoic Acids	5
Fats	5
Hydroxybutyrates	5
Inositol	5

Leukotrienes	5
Linoleic Acid	5
Margarine	5
Neodymium	5
Phosphatidylinositol 4,5-Diphosphate	5
Riboflavin	5
Sodium Isotopes	5
SRS-A	5
Strontium Radioisotopes	5
Argon	4
Bromides	4
Deuterium	4
Fat Emulsions, Intravenous	4
Folic Acid	4
Hydroxyeicosatetraenoic Acids	4
Iodides	4
Iodized Oil	4
Niacin	4
Phosphatidylinositol Phosphates	4
Silicon	4
Sulfur	4
Thiamine	4
Valerates	4
alpha-Tocopherol	3
Aluminum	3
Caproates	3
Carboprost	3
Cerium	3
Ceroid	3
Cesium Isotopes	3
Chromium	3
Gangliosides	3
Glycerophosphates	3
Isotopes	3
Lanosterol	3
Linseed Oil	3
Lipoproteins, LDL Cholesterol	3
Niacinamide	3
Pentanols	3
Radioactive Tracers	3
Silver	3
Thiamine Pyrophosphate	3
Tin	3
4-Aminobenzoic Acid	2
Butanols	2
Calcitriol	2
Calcium Isotopes	2
Castor Oil	2
Cesium Radioisotopes	2
Cholesterol Esters	2
Diphosphates	2

Ergocalciferols	2
Gold Isotopes	2
Iron Isotopes	2
Iron Radioisotopes	2
Lead	2
Leukotriene B4	2
Lipoproteins, VLDL Cholesterol	2
Mercury Isotopes	2
Nitrites	2
Oils, Volatile	2
Oxides	2
Prostaglandins A	2
Safflower Oil	2
Selenium	2
Sphingolipids	2
Sterols	2
Superoxides	2
Thiosulfates	2
Trace Elements	2
Turpentine	2
Uranium	2
Ytterbium	2
Yttrium	2
1-Butanol	1
Acetic Acid	1
alpha-Linolenic Acid	1
Antimony	1
Arsenic	1
Beryllium	1
Biotin	1
Borates	1
Caprylates	1
Carnitine	1
Cations	1
Cations, Monovalent	1
Cerebrosides	1
Cerium Isotopes	1
Cerium Radioisotopes	1
Cesium	1
Chlorates	1
Cottonseed Oil	1
Cystine	1
Decanoic Acids	1
Dysprosium	1
Elements	1
Erbium	1
Fatty Acids, Monounsaturated	1
Fatty Alcohols	1
Ferrocyanides	1
Gadolinium	1
gamma-Linolenic Acid	1
Germanium	1
Glycerylphosphoryl-choline	1

Glycolipids	1
Gold Radioisotopes	1
Graphite	1
Helium	1
Hexanols	1
Holmium	1
Hydroxycholesterols	1
Lauric Acids	1
Leukotriene A4	1
Lipid A	1
Lipid Bilayers	1
Lipofuscin	1
Lutetium	1
Molybdenum	1
Osmium	1
Oxygen Isotopes	1
Palladium	1
Palmitic Acid	1
Pantothenic Acid	1
Phosphatidylglycerols	1
Plant Oils	1
Plasmalogens	1
Platinum	1
Praseodymium	1

Prostaglandins A, Synthetic	1
Protons	1
Pyridoxal	1
Radium	1
Radon	1
Ruthenium	1
Samarium	1
Stearates	1
Sulfites	1
Sulfur Isotopes	1
Sulfur Radioisotopes	1
Tantalum	1
Terbium	1
Thallium	1
Titanium	1
Triolein	1
Tungsten	1
Valproic Acid	1
Vitamin K 3	1
Waxes	1
Yttrium Isotopes	1

Résumé

L´information disponible dans les bases de données bibliographiques est une information datée, validée par un processus long qui la rend peu innovante. Dans leur mode d´exploitation, les bases de données bibliographiques sont classiquement interrogées de manière booléenne. Le résultat d´une requête est donc un ensemble d´informations connues qui n´apporte en lui-même aucune nouveauté.
Pourtant, en 1985, Don Swanson propose une méthode originale pour extraire de bases de donnés une information innovante. Son raisonnement est basé sur une exploitation systématique de la littérature biomédicale afin de dégager des connexions latentes entre différentes connaissances bien établies. Ses travaux montrent le potentiel insoupçonné des bases bibliographiques dans la révélation et la découverte de connaissances. Cet intérêt ne tient pas tant à la nature de l´information disponible qu´à la méthodologie utilisée. Cette méthodologie générale s´applique de façon privilégiée dans un environnement d´information validée et structurée ce qui est le cas de l´information bibliographique. Nous proposons de tester la robustesse de la théorie de Swanson en présentant les méthodes qu´elle a inspirées et qui conduisent toutes aux mêmes conclusions. Nous exposons ensuite, comment à partir de sources d´information biomédicales publiques, nous avons développé un système de découverte de connaissances basé sur la littérature.

Mots-Clés : Découverte de Connaissances, Bases de Données Bibliographiques, Bibliométrie.

Abstract

The information available in bibliographic databases is dated and validated by a long process and becomes not very innovative. Usually bibliographic databases are consultated in a boolean way. The result of a request represente is a set of known which do not bring any additional novelty.
In 1985 Don Swanson proposed an original method to draw out innovative information from bibliographic databases. His reasoning is based on systematic use of the biomedical literature to draw the latent connections between different well established knowledges. He demonstrated unsuspected potential of bibliographic databases in knowledge discovery. The value of his work did not lie in the nature of the available information but consisted in the methodology he used. This general methodology was mainly applied on validated and structured information that is bibliographic information. We propose to test the robustness of Swanson's theory by setting out the methods inspired by this theory. These methods led to the same conclusions as Don Swanson's ones. Then we explain how we developed a knowledge discovery system based on the literature available from public biomedical information sources.

Key-Words : Knowledge Discovery, Bibliographic Databases, Bibliometry.