



HAL
open science

Découverte de connaissances dans les bases de données bibliographiques. Modèles expérimentaux autour de la première expérience de Swanson

Jean-Dominique Pierret, Fabrizio Dolfi, Luc Quoniam, Eric Boutin, Edson Luiz Riccio

► To cite this version:

Jean-Dominique Pierret, Fabrizio Dolfi, Luc Quoniam, Eric Boutin, Edson Luiz Riccio. Découverte de connaissances dans les bases de données bibliographiques. Modèles expérimentaux autour de la première expérience de Swanson. *Informations, Savoirs, Décisions et Médiations [Informations, Sciences for Decisions Making]*, 2005, 20 (244). sic_00001383

HAL Id: sic_00001383

https://archivesic.ccsd.cnrs.fr/sic_00001383v1

Submitted on 23 Mar 2005

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

**DECOUVERTE DE CONNAISSANCES DANS LES BASES DE DONNEES
BIBLIOGRAPHIQUES
MODELES EXPERIMENTAUX AUTOUR DE LA PREMIERE HYPOTHESE DE SWANSON**

Jean-Dominique Pierret,

Galderma R&D

jeandominique.pierret@galderma.com + 33 4 93 95 70 48

Fabrizio Dolfi

Galderma R&D

fabrizio.dolfi@galderma.com + 33 4 93 95 70 53

Luc Quoniam

Laboratoire LePont

quoniam@univ-tln.fr + 33 4 94 14 23 56

Eric Boutin

Laboratoire LePont

boutin@univ-tln.fr + 33 4 94 14 23 56

Edson Luiz Riccio

Laboratorio de Tecnologia e Sistemas de Informaçao

elriccio@usp.br (011) 3091-5820

Adresse professionnelle

Galderma R&D ★ 635, route des Lucioles ★ BP 87 ★ F-06902 Sophia-Antipolis Cedex
Laboratoire LePont ★ Université de Toulon-Var ★ BP 132 ★ F-83957 La Garde Cedex
TECSI ★ Av. Prof. Luciano Gualberto, 908 FEA 3. Cidade Universitária São Paulo ★ Brésil

Résumé : En 1985, Don Swanson propose une hypothèse originale pour traiter la maladie de Raynaud avec de l'huile de poisson. Son raisonnement est basé sur une exploitation systématique de la littérature biomédicale afin de dégager des connexions latentes entre différentes connaissances bien établies. Aujourd'hui, sa première découverte et sa théorie de *knowledge discovery in databases* ont été reprises par différentes équipes pour tester des systèmes informatisés d'aide à la génération d'hypothèses.

Cet article reprend deux exemples d'utilisation de la théorie de Swanson et développe une application originale basée sur l'exploitation d'un thésaurus biomédical.

Mots clés : découverte de connaissances, bases de données bibliographiques.

Key words: knowledge discovery, bibliographic databases.

Découverte de connaissances dans les bases de données bibliographiques

Modèles expérimentaux autour de la première hypothèse de Swanson

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 (Knowledge Discovery in Databases – KDD) [Swanson, 1986]. Avec l'explosion du nombre de publications scientifiques et la fragmentation des communautés de chercheurs autour de thèmes toujours plus complexes, il existe certainement des connexions 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 littérature AB et BC, ce lien restera latent. Son travail sur la maladie de Raynaud¹ 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 effets, 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 de la maladie de Raynaud. Après analyse de la littérature, Swanson fut le premier à proposer d'utiliser l'huile de poisson comme traitement de la maladie de Raynaud. Notre article précédant [Pierret, 2004] reprend et détaille le modèle de Swanson : ainsi, 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 implicite AC ne soit pas connue. Swanson a montré à plusieurs reprises [Swanson, 1986, 1988, 1990] que des parties disjointes de la connaissance biomédicale peuvent être connectées en étudiant leurs littératures respectives selon ce modèle transitif où :

¹ 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.

- 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 :

- $A \rightarrow B \rightarrow C$, 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* ($C \rightarrow B \rightarrow A$ est bien entendu inclus dans ce cas).
- $A \rightarrow B \leftarrow C$, qui va explorer les différentes relations B possibles entre A et C, processus clos permettant de *tester une hypothèse*.

Par la suite, Swanson, avec Neil Smalheiser, ont développé Arrowsmith, un système informatique, d'aide à l'exploration d'hypothèses, à partir de la littérature. D'autres équipes se sont intéressées au travail de Swanson et ont adapté sa méthode. Nous allons exposer dans la suite de cet article deux travaux réalisés à partir de l'exemple de la maladie de Raynaud ainsi que les résultats que nous avons obtenu en utilisant le thésaurus de la NLM², le MeSH³.

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 Medline⁴. 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

² National Library of Medicine, Bethesda, Maryland.

³ MeSH, Medical Subject Heading, voir § 3.

⁴ Base de données bibliographique biomédicale produite par la NLM, accessible par l'interface PubMed : <http://www.ncbi.nlm.nih.gov>.

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 télécharger 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 Arrowsmith pour vérifier directement 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 télécharger 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.

3 – A partir des termes B conservés, Arrowsmith permet, pour un terme B_n donné, d'éditer la liste des titres AB_n juxtaposés à ceux citant B_nC 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-grams). Relevons au passage qu'il utilise une technique de recherche non-booleenne, 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. Pour 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, 1998] et [Swanson, 1999]:

- maladie de Raynaud et huile de poisson,
- migraine et magnésium : une déficience en magnésium comme facteur aggravant de la migraine,
- indométhacine et maladie d'Alzheimer : effets indésirables possibles de l'indométhacine 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'Arrowsmith⁵ qui, sur le fond sont identiques et sont basées sur un processus de découverte de connaissance clos. 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 maintien 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).

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'UMLS⁶ 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, thésaurii, classifications), tel que le MeSH ou le MeDRA (*Medical Dictionary for Regulatory*

⁵ Arrowsmith 3.0 (kiwi.uchicago.edu) et Arrowsmith "UIC" (arrowsmith.psych.uic.edu/arrowsmith_uic/index.html)

⁶ UMLS : Unified Medical Language System (www.nlm.nih.gov/research/umls/)

Activities Terminology) en différentes langues, ou encore Gene Ontology⁷ ou HUGO Gene Nomenclature⁸. UMLS est organisé en trois parties :

- le méta thésaurus, 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 la première hypothèse de Swanson [Weeber 2000]. 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.

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 1246 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 deformability, blood viscosity, platelet adhesiveness* et *hemorheology*. Ces termes sont sélectionnés comme concepts B.

2.2 – Générer B→A

Partant de ces concepts B, le DAD décharge 10611 références bibliographiques de PubMed. Les phrases où apparaissent les concepts B sont analysées pour extraire les concepts associées : 7702 concepts sont identifiés parmi lesquels 6747 n'apparaissent pas dans les phrases extraites de la littérature C. Ne sont alors retenus 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

⁷ Gene Ontology : www.geneontology.org

⁸ HUGO : www.gene.ucl.ac.uk/nomenclature

(*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, retenant l'huile de poisson comme concept A.

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, 1986]. 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.

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*.

3 – UTILISATION DU MeSH DANS UN PROCESSUS DE DECOUVERTE DE CONNAISSANCE

Dans la suite de cet article, les descripteurs du MeSH seront écrits en utilisant la police de caractère `courrier new`.

Le MeSH⁹ est le thésaurus 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

⁹ MeSH home page : www.nlm.nih.gov/mesh

en ligne à partir de plusieurs systèmes différents (la MeSH Database¹⁰ de PubMed ou le MeSH Browser¹¹). Mis à jour chaque année, la version 2004 comporte plus de 22000 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 comme par exemple : *Anatomy, Organisms, Diseases* ou *Chemicals and Drugs*, mais aussi *Humanities* ou *Information Science*. Il est ainsi facile d'isoler les descripteurs relatifs aux maladies, à la physiologie ou aux médicaments et substances chimiques.

Les références bibliographiques de Medline sont indexées avec le MeSH. Notre propos est de montrer qu'il est possible de simuler l'expérience de Swanson à partir du MeSH. Ainsi, l'interrogation de Medline sur la maladie de Raynaud donne environ un millier de références bibliographiques (jusqu'en novembre 1985). Le traitement de ces références avec un logiciel de reformatage et un système de gestion de base de données relationnel (SGBDr), combiné à l'utilisation de la partie physiologie du MeSH, permet de connaître les concepts relatifs à la physiologie les plus fréquemment utilisés pour indexer la littérature relative à la maladie de Raynaud. Il en est de même avec la littérature relative au phénomènes physiologiques impliqués dans la maladie de Raynaud : la partie du MeSH correspondant aux médicaments et molécules couplée à l'utilisation du logiciel de reformatage et du SGBDr nous conduit à identifier les substances les plus fréquemment citées dans ce contexte.

Nous travaillons en suivant l'hypothèse selon laquelle lorsque deux descripteurs apparaissent ensembles dans une même référence bibliographique, ils sont alors très certainement liés. Ce lien est d'autant plus probable si sur un corpus, deux descripteurs apparaissent fréquemment ensembles. 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.

3.1 Matériel et méthode

Le MeSH Tree est déchargé à partir du site de la NLM¹⁰ et trois fichiers de descripteurs sont préparés :

1. descripteurs relatifs à la physiologie,
2. descripteurs relatifs aux vitamines, lipides, ions, éléments et isotopes,

¹⁰ MeSH Database :

www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=mesh

¹¹ MeSH Browser :

www.nlm.nih.gov/mesh/2005/MBrowser.html

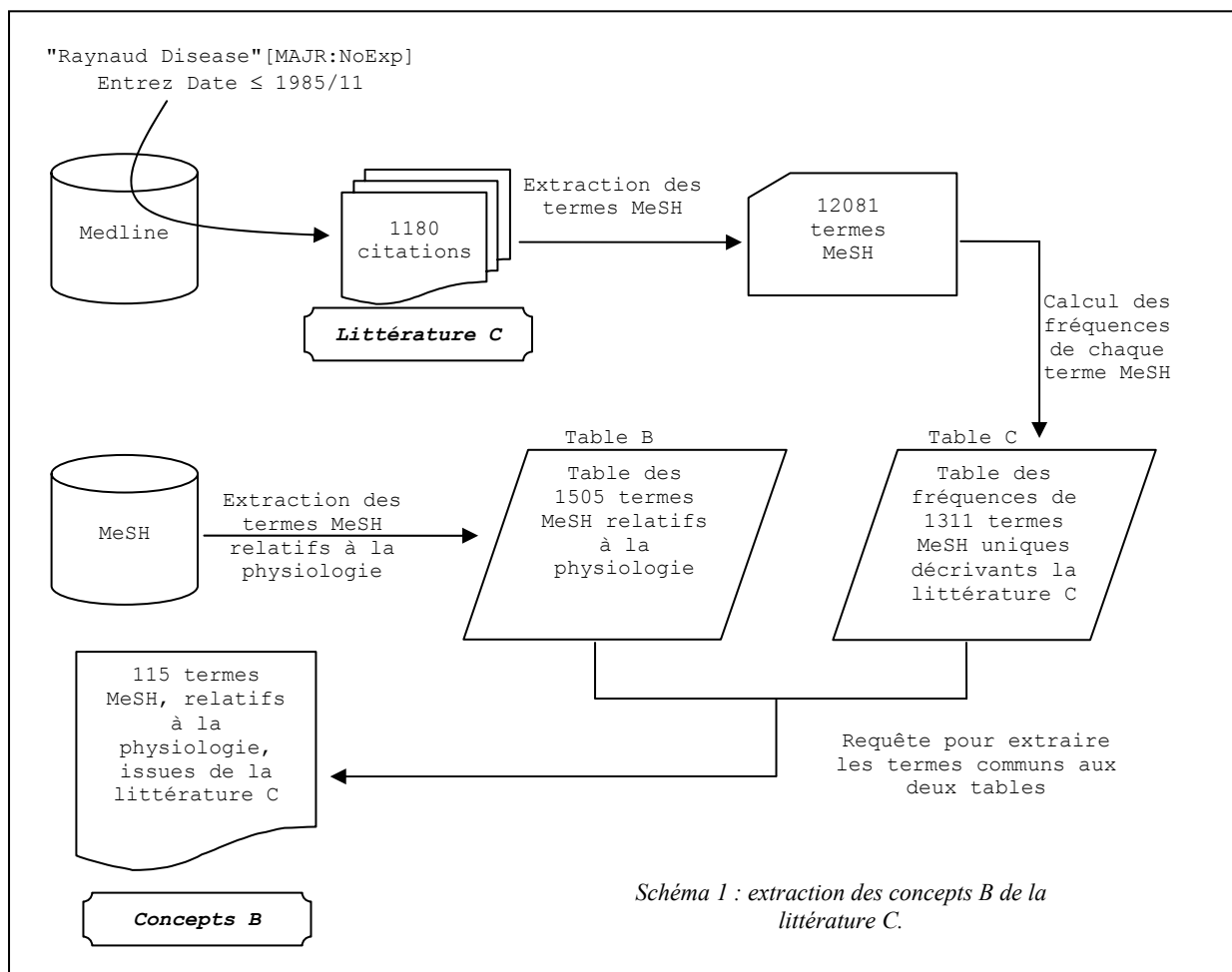
3. le dernier fichier englobe le second, et contient les descripteurs relatifs aux médicaments et substance chimiques.

Le traitement des données est réalisé avec un logiciel de reformatage et un système de bases de données relationnel. Le logiciel de reformatage est employé pour éditer les fichiers bibliographiques et les épurer afin de ne conserver que les descripteurs et éliminer tous les autres champs des références. Ainsi traité, le fichier bibliographique, devenu une liste de descripteurs est chargé dans une table du SGBDr. Il permet alors de calculer la fréquence des descripteurs dans cette table et de la dédoubler avec une autre table qui contient, par exemple, la partie physiologie du MeSH. Le dédoublonnage est la fonction pivot qui sert de "filtre sémantique" (Schéma 1).

Ainsi en déchargeant de PubMed les références sur la maladie de Raynaud (littérature C) et en isolant

par reformatage les descripteurs de ces références, le SGBDr va permettre de stocker cette liste descripteurs/fréquences dans un table C. Si une autre table B contient une liste de descripteurs relatifs à la physiologie, dédoubler les tables B et C sur les noms des descripteurs revient à extraire les concepts B de la littérature C. En identifiant, parmi les descripteurs B issus de la littérature C, un concept intéressant (agrégation plaquettaire par exemple), il faut de nouveau interroger PubMed et télécharger les références relatives à ce concept, c'est la littérature B. Toujours selon le même mode opératoire, on extrait alors de cette littérature les descripteurs relatifs aux vitamines, lipides, ..., les concepts A. L'huile de poisson et ses composés figurent en bonne position.

PubMed est interrogé en utilisant comme limite temporelle supérieure novembre 1985, afin d'employer un corpus bibliographique similaire à celui utilisé par Swanson.



3.2 Résultats

L'ensemble des interrogations de PubMed a été réalisé le 18 décembre 2004. Pour les étapes de sélection des littératures C, B et A, nous avons choisi d'interroger avec les descripteurs majeurs, c'est-à-dire ceux qui, pour une notice donnée, représentent le cœur du propos de l'article indexé. Nous augmentons ainsi la pertinence des liens que nous mettrons en évidence. Le détail des requêtes figurent en notes de bas de page.

La littérature C, relative à la maladie de Raynaud comprend 1180 citations¹². L'extraction des concepts relatifs à la physiologie donne une liste de 117 descripteurs (majeurs et simples). Les plus fréquents sont (les occurrences figurent entre parenthèses) :

- 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) etc ...

Cette liste montre de manière évidente que la physiopathologie de la maladie de Raynaud est liée, entre autre, à des problèmes de circulation sanguine périphérique, à l'hémodynamique ou à la coagulation. Reprenant le raisonnement de Swanson en analysant cette liste, nous isolons les cinq concepts B suivants :

- Blood Viscosity (41)
- Blood Flow Velocity (32)
- Platelet Aggregation (8)
- Platelet Adhesiveness (3)
- Blood Coagulation (6)

Ces concepts B sont identiques à ceux utilisés par Weeber pour simuler l'expérience de Swanson [Weeber, 2000].

L'interrogation de PubMed à partir des concepts B que nous avons sélectionnés donne 16393 références¹³. Nous isolons de cette littérature B 308 descripteurs relatifs aux vitamines, lipides, ions, éléments et isotopes (les *diet factors* de Swanson). Ce sont les concepts A. Les prostaglandines figurent en tête de liste :

- Arachidonic Acids (516)
- Epoprostenol (467)

¹² "Raynaud Disease"[MAJR:NoExp] Limits : Entrez Date to 1985/11

¹³ "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

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 51^{ème} position :

- Eicosapentaenoic Acid (35)
- Fishes 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.

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, 1986].

Cependant une autre voie est envisageable, pour aider à identifier plus rapidement les articles complémentaires pertinents. Il s'agit de tester l'hypothèse C→B←A en recherchant parmi les concepts relatifs à la physiologie, lesquels sont communs aux littératures A (*fish oils*) et C (*Raynaud disease*). La littérature sur l'huile de poisson est récupérée à partir de PubMed¹⁴. Cela donne 212 références bibliographiques, desquelles sont ensuite extraits les descripteurs relatifs à la physiologie. Enfin Access permet de rechercher les concepts physiologiques communs à la littérature sur l'huile de poisson et à la littérature sur la maladie de Raynaud. Le tableau 1 est édité sous Excel à partir de requêtes Access. Il contient tous les descripteurs MeSH relatifs à la physiologie, commun aux littératures sur l'huile de poisson, 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

¹⁴ "Fish Oils"[MAJR] Limits: Entrez Date to 1985/11

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 également 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.

Tableau 1

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

3.3 Au delà de l'huile de poisson

Restant dans les mêmes conditions, c'est à dire en utilisant les citations ajoutées à Medline au plus tard en novembre 1985, nous avons traité les 16393 références bibliographiques de la littérature B (voir paragraphe 3.2 ci-dessus¹³), en extrayant les concepts relatifs aux médicaments et substances chimiques. Nous obtenons 1984 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 (1752 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 apportée par Gachet en 1992 et Fredholm précisera la nature des différents récepteurs à l'adénosine ou récepteurs purinergique 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 USPTO¹⁵ 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 Pharmaprojects¹⁶ 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 avons

¹⁵ United States Patent and Trademark Office, patents.uspto.gov.

¹⁶ Pharmaprojects, base de données éditée par PJB Publications rapportant le status de candidats médicaments en développement.

formulé à posteriori cette hypothèse, avec des données antérieures 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.

3.3 Commentaires

La répétition de la première expérience de Swanson en utilisant le MeSH nous a montré qu'en recherchant des connexions au niveau des références bibliographique 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 utilisée, des champs retenus pour établir les connexions témoignent 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 connexions qui prennent sens 15 à 20 ans plus tard.

4 – CONCLUSION

Don Swanson a largement détaillé le mode opératoire de sa découverte, basé sur une utilisation classique de bases de données bibliographiques (Medline ou SciSearch¹⁷) 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 [Srinivasan, 2004], [Gordon, 1996], [Swanson, 1999], [Weeber, 2000]. Les résultats obtenus permettent toujours de proposer l'utilisation de l'huile de poisson comme traitement. L'hypothèse qu'il a proposé n'était pas si évidente qu'il n'y parait à posteriori – sinon, d'autres l'auraient formulé 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 formuler 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 ...), 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 thésaurii associés avant toute utilisation du logiciel, bien connaître le sujet sur lequel ils travaillent, faire preuve d'ingéniosité et être capable d'identifier des connexions 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 complémentaire, 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, la méthode basée sur le MeSH permet de générer rapidement la liste des termes B. Elle 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

¹⁷ SciSearch, aujourd'hui édité par Thomson et publié par l'ISI (Institute for Scientific Information), est une base de données bibliographique dont la particularité est de donner accès, pour chaque référence, aux articles cités. Cet outil a permis à Swanson de vérifier si les corpus sur la maladie de Raynaud et sur l'huile de poisson citent ou non les mêmes articles et donc de « mesurer » la séparation entre ces deux littératures.

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 bio informatiques) 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 NCBI¹⁸). Le *knowledge discovery in databases* n'est pas loin.

Quel que soit l'outil utilisé – Arrowsmith, le DAD, le MeSH – le schéma de Swanson et les techniques associées sont des aides précieuses à l'exploitation des données bibliographiques. Medline contient aujourd'hui plus de 12 millions de citations, ce qui donne un nombre formidable de paires d'articles. 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 la formidable quantité 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 intégrées dans un processus de recherche et développement, n'attendant que la confrontation aux résultats expérimentaux.

¹⁸ National Center for Biotechnology Information : www.ncbi.nlm.nih.gov/Tools/

BIBLIOGRAPHIE

- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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°109.
- 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 Pharmacology*. Vol. 259, n°1, p. 43-49.
- Shryock, J.C., Belardinelli, L. (1997), "Adenosine and adenosine receptors in the cardiovascular system: biochemistry, physiology, and pharmacology", *American Journal of Cardiology*. Vol.79, n°12A, p. 2-10.
- Smalheiser, N.R., Swanson D.R. (1998), "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., 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.
- Srinivasan, P. (2004), "Text mining: generating hypotheses from MEDLINE", *Journal of the American Society for Information Science*. Vol. 55, n°5, p. 396-413.
- Swanson, D.R. (1986), "Fish oil, Raynaud's syndrome, and undiscovered public knowledge", *Perspectives in Biology and Medicine*. Vol. 30, n°1, p. 7-18.
- 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. (1990), "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., 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.
- Tabrizchi, R., Bedi, S. (2001), "Pharmacology of adenosine receptors in the vasculature", *Pharmacology and Therapeutics*. Vol. 91, n°2 p. 133-147.
- Weeber, M.A., 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.

Annexe

Exemple d'une citation Medline issue de PubMed

Les commentaires sont en *italique/gras*.

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

Word est employé pour isoler le champ suivant (multivaleurs) et ne conserver que les descripteurs. Par exemple, pour la première ligne de "MH", il faut extraire le terme MeSH "Abstracting and Indexing" et éliminer l'intitulé du champ ("MH - ") ainsi que le qualificatif "/methods" (subheading). Lorsque le terme MeSH est un descripteur majeur, l'étoile * doit être également éliminée.

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.