



**BIOINFORMATIQUE ET
CLINICOGÉNOMIQUE AU SERVICE DE
L'INNOVATION THÉRAPEUTIQUE**

Compte-rendu de la journée du
27 novembre 2003



Sommaire

PROPOS INTRODUCTIFS	3	Guillaume LIÈVRE, <i>Directeur Général, IT.OMICS</i>	51
Etienne VERVAECKE, <i>Directeur du GIE Eurasanté</i>		Vincent FERT, <i>CEO - Ipsogen</i>	57
Pr Philippe AMOUYEL, <i>Directeur de la Génopole de Lille Nord Pas-de-Calais et de l'Institut Pasteur de Lille</i>		Clarisse DHAENENS, <i>Laboratoire d'Informatique Fondamentale de Lille, Université de Lille I</i>	63
		Discussion	69
ORGANISATION ET FÉDÉRATION DES DONNÉES PRODUITES PAR LA CLINICOGÉNOMIQUE	13		
Pr Régis BEUSCART, <i>Directeur du Centre de Recherche en Informatique Médicale de Lille</i>	15		
Dr Christophe ROUMIER, <i>Laboratoire d'Hématologie du CHRU de Lille - INSERM U524</i>	19		
Teresa GARCIA, <i>Responsable de l'identification des cibles, société Proskelia</i>	23		
Pascal SEMPÉ, <i>Business Development Executive, IBM Division Sciences du Vivant</i>	31		
Thierry PEUCHOT, <i>IBM Division Sciences du Vivant</i>	39		
Discussion	43		
CLINICOGÉNOMIQUE, VERS DE NOUVELLES MÉTHODES BIOINFORMATIQUES D'INTERPRÉTATION DES DONNÉES CLINIQUES	45		
Dr Geneviève Martin, <i>Chef de projet et coordinatrice Nouveaux Développements, Genfit</i>	47		

PROPOS INTRODUCTIFS



EURASANTÉ

Parc Eurasanté - 310 av. Eugène Avinée - 59120 Loos

Tél. 03 28 55 90 60

contact@eurasante.com

www.eurasante.com

Etienne Vervaecke

Directeur du GIE Eurasanté

Eurasanté est l'agence de développement économique de la région Nord-Pas-de-Calais dédiée aux activités liées à la santé, à la médecine et aux biotechnologies.

Nous avons pour mission le développement de ce pôle économique et scientifique, avec comme activités principales :

- le soutien au développement et à la création d'entreprises dans les secteurs de la biologie, de la santé et de la médecine
- le soutien aux projets de création d'entreprises innovantes dans le cadre du bioincubateur Eurasanté, soutenu par les collectivités locales, la Région et le Ministère de la Recherche
- l'accompagnement à l'implantation dans la région d'entreprises extra régionales et étrangères
- la valorisation économique des savoir-faire cliniques et de recherche des professionnels de santé du Nord-Pas-de-Calais

La bioinformatique est un secteur que nous croyons être prometteur en terme de création d'emplois et de valeur dans notre région. La société IT.Omics est un bon exemple de réussite sur ce créneau. C'est la raison pour laquelle nous avons souhaité nous associer avec elle, ainsi qu'avec la division Sciences du Vivant d'IBM pour organiser cette manifestation autour du rôle de la bioinformatique dans

cette source d'innovation thérapeutique qu'est le croisement entre la clinique et la génomique.

Cette conférence permettra de détailler la nature des outils bioinformatiques au sein de la génopole et des plateaux hospitalo-universitaires de notre région. Je vais laisser la parole au professeur Amouyel que nous remercions de nous accueillir au sein de son Institut et qui nous fait le plaisir d'inaugurer cette manifestation.



INSTITUT PASTEUR DE LILLE

1, rue du Pr. Calmette - BP 245 - 59019 Lille Cedex

Tél. 03 20 87 73 16

philippe.amouyel@pasteur-lille.fr

www.pasteur-lille.fr

Pr. Philippe Amouyel
Directeur de la Génopole de
Lille Nord Pas-de-Calais et
de l'Institut Pasteur de Lille

La mission de l'Institut Pasteur de Lille comporte quatre grands axes :

- la recherche en biologie et en santé
- l'expertise et la formation en environnement et santé
- la prévention et la santé publique, notamment dans le cadre de la vaccination
- l'innovation, le transfert et la valorisation technologiques

Sur le plan de la recherche, nous développons essentiellement la médecine moléculaire. L'objectif de l'Institut Pasteur n'est pas de faire du soin. Notre rôle s'arrête dans ce domaine aux actions de conseil et de prévention.

Notre objectif essentiel est d'essayer d'anticiper, d'intégrer et de maîtriser les nouvelles technologies de la biologie qui vont dessiner la santé et la médecine de demain.

Le campus regroupe actuellement 17 équipes, 629 chercheurs, ingénieurs et techniciens et 146 doctorants et DEA.

Notre recherche couvre de nombreux champs, des maladies cardiovasculaires à la pharmacologie moléculaire en passant par les vaccins et la nutrition, qui partagent une vision commune : l'utilisation intensive

des outils les plus performants et notamment la génomique, la protéomique et la transcriptomique (voir figures 1 et 2).

Ces technologies ont explosé au cours des 5 à 10 dernières années et ont entraîné la production et l'accumulation d'un volume considérable d'informations. Il y a encore 5 ans en France et 10 ans dans certains pays anglo-saxons, on avait l'habitude de travailler de manière « linéaire », en étudiant un gène, une fonction ou une protéine à la fois. Aujourd'hui, les recherches changent d'échelle et concernent l'observation concomitante de plusieurs gènes en prenant en compte à la fois leur structure, leur fonctionnement et leurs interactions.

Ce changement d'échelle s'observe par exemple au niveau des études d'épidémiologie, où la taille des échantillons est passée d'une centaine d'individus à des séries statistiques minimales de 1000 voire 10000 sujets.

Prenons un autre exemple, celui d'une étude prospective française visant à étudier les relations entre la démence et les maladies cardiovasculaires. Pour obtenir des résultats crédibles en un laps de temps raisonnable, (3 à 6 ans de surveillance), il a fallu recruter environ 10000 sujets de plus de 65 ans ce qui représente un nombre considérable d'informations : 9684 questionnaires incluant des informations extrêmement

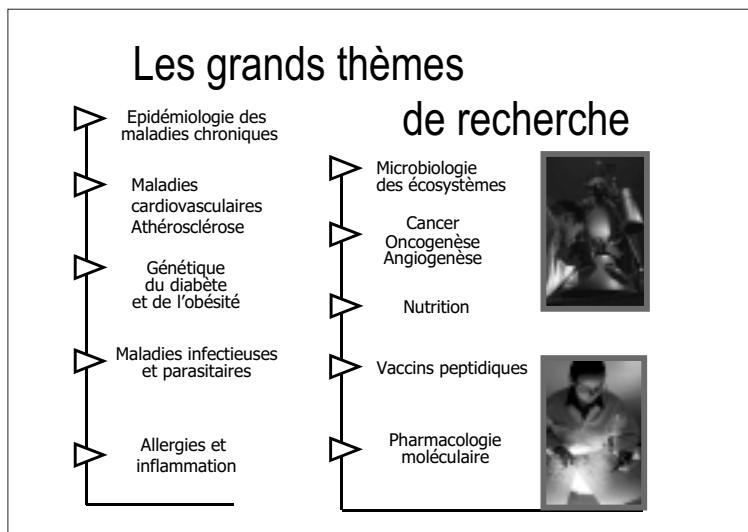


Figure 1

8

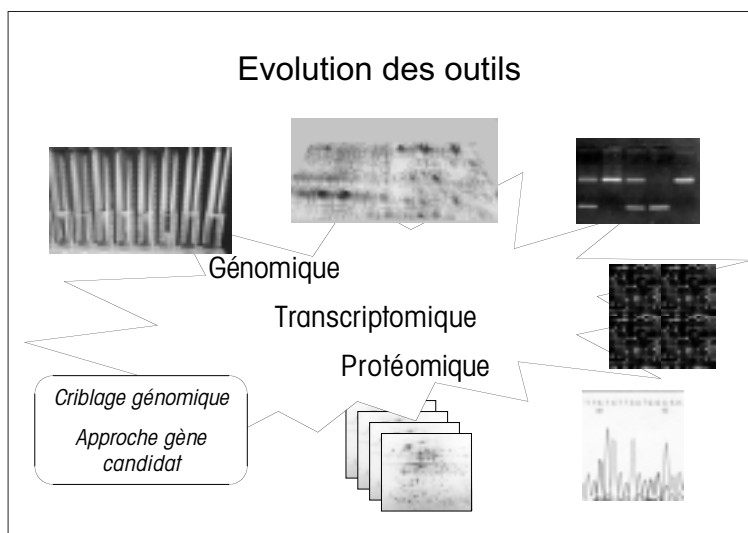


Figure 2

détaillées comme l'imagerie cérébrale (6000 IRM sont prévues dans cette étude) (voir figure 3).

C'est la plus grande étude européenne

211126 microtubes. Sa constitution et la gestion des échantillons nécessitent le traitement d'un très grand nombre de données compte tenu du nombre d'individus, des événements analysés et des

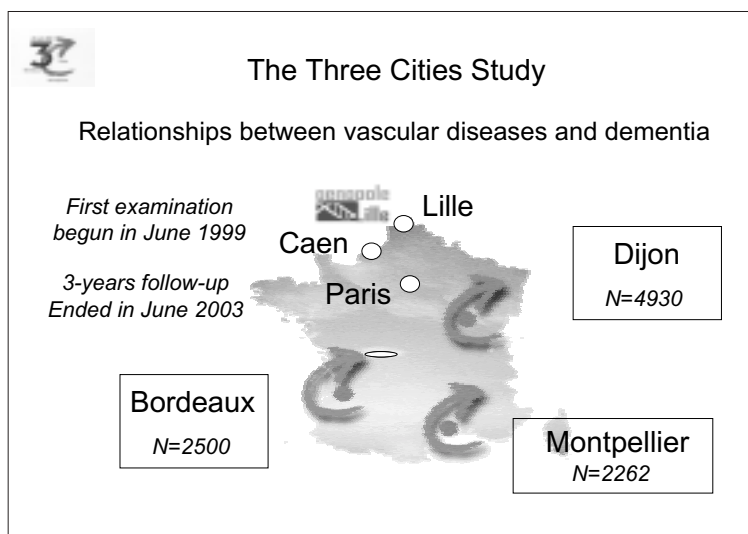


Figure 3

incluant des informations venant de l'IRM corrélées à des questionnaires généraux, aux fonctions cognitives, à des échantillons biologiques, à l'échographie carotidienne et à l'électrocardiogramme.

Aujourd'hui, environ 3000 IRM ont été traitées. Ces données représentent 196 gigaoctets d'informations dont l'exploitation nécessite une grande puissance de calcul.

Une banque d'ADN de tissus de tous types a été développée qui contient

mesures biologiques (voir figure 4).

Nous commençons à disposer d'outils pour gérer ce type de volumes d'échantillons. Dans le cadre de la Génopole de Lille, nous avons pu nous doter d'une plate-forme de génomique à haut débit qui peut traiter jusqu'à 32 méga bases et 900 000 génotypes par mois (voir figure 5).

Nous avons développé le même type d'approche, dans le cadre d'un programme plus vaste, sur la recherche de nouveaux facteurs de susceptibilité.

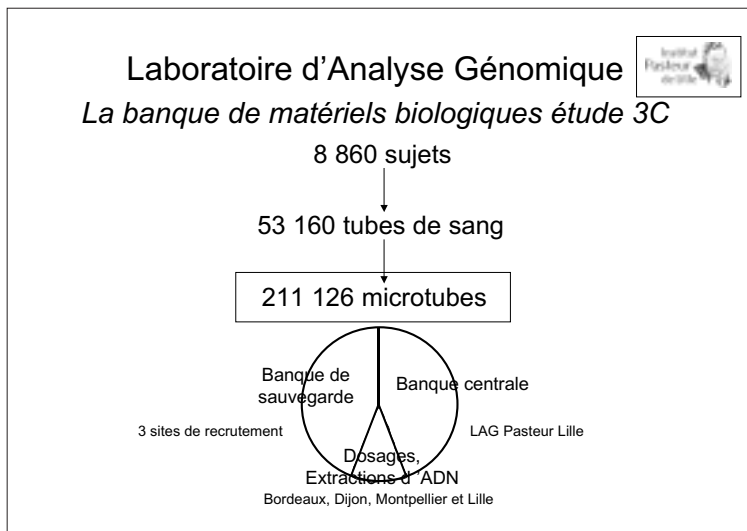




Figure 4

10


High throughput genomics




Tecan Freedom robot, 1600 plates 384/month




2 waves dHPLC, 3 000 samples/month



PCR, > 400 000 ampli. / month



Applera 3730 XL, 32 megabases/month



Applera Taqman 7900, 900 000 genotypes/month

1 CEO, 1 post-PhD, 1 PhD, 1 engineer, 2 engineering assistants, 1 commercial engineer, 1 bioinformatician, 1 administration assistant + scientific board

Figure 5

Certaines approches transcriptomiques sont aussi mises en œuvre. Une trentaine d'échantillons de cerveaux, qui ont permis de produire 18 puces de type Affymetrix, sont actuellement en cours d'analyse, à la recherche de nouveaux facteurs de susceptibilité à la maladie d'Alzheimer.

Nous développons aussi des études différentielles à partir du protéome. Cependant, le terme de high-throughput (haut débit) n'est pas encore d'actualité pour la protéomique. Un certain nombre d'éléments technologiques limitent la vitesse de production des informations. On arrive néanmoins à traiter entre 200 et 300 spots différents sur quelques mois, une fréquence assez rapide à partir du moment où l'on identifie cette différenciation.

En conclusion, nous disposons d'outils qui sont capables de produire plus d'informations qu'un seul chercheur dans son laboratoire. Il est donc nécessaire de définir une stratégie d'analyse de ces informations.

Les approches de haut débit permettent de récolter et de qualifier des informations sans connaître les processus physiologiques qui leur sont sous-jacents. Elles ont permis d'établir que les éléments biologiques à l'origine de la vie étaient en nombre plus ou moins finis et que l'on pouvait les maîtriser.

Plus que dans d'autres pays où les investissements sur ces champs sont plus lourds, nous avons besoin d'une véritable révolution chez les chercheurs français. Pour maîtriser ces outils qui seront un facteur essentiel de l'innovation au service de la santé des populations et pour pouvoir assurer les analyses et les interprétations issues des volumes gigantesques d'infor-

mations générés par ces outils. C'est là que la bioinformatique a son rôle à jouer. Il est important que dès à présent, les chercheurs français prennent conscience de l'importance de ces nouvelles technologies et que les laboratoires s'approprient au plus vite ces nouvelles approches de gestion des informations scientifiques dont les flux ne feront qu'augmenter.

ORGANISATION ET
FÉDÉRATION DES DONNÉES
PRODUITES PAR LA
CLINICO-GÉNOMIQUE

Pr. Régis Beuscart
Directeur du Centre de Recherche
en Informatique Médicale
de Lille

CERIM
Faculté de Médecine Henri Warembourg
Avenue Eugène Avinée - 59045 Lille Cedex
Tél. 03 20 44 60 42 - rbeuscart@chru-lille.fr

Outre mes activités d'enseignement et de recherche au centre d'Études et de Recherche d'Informatique Médicale de la Faculté de Médecine de Lille, je suis également responsable du système d'information du CHRU de Lille qui dispose d'une gestion du dossier médical partagé entre tous les centres de production de soins représentant 1,5 million de patients, sur une population régionale totale de 4 millions d'habitants.

La biologie génère de plus en plus de données, de façon complètement automatisée, tout comme la génomique, la transcriptomique et la protéomique. L'acquisition de ces données devient un facteur essentiel de la compréhension du vivant et face à cela, nous avons besoin d'outils qui permettent de résoudre les problèmes suivants :

- cataloguer les données en construisant des bases dans la perspective de leur exploitation ultérieure : une base mal conçue rendra difficile l'exploitation des données qu'elle contient.
- organiser ces bases en limitant leur dispersion,
- les exploiter pour permettre d'inférer de nouvelles connaissances.

La bioinformatique regroupe généralement sept thèmes. D'une part, il s'agit de bien connaître la nature des données biologiques car celle-ci diffère des informations que les médecins et les informaticiens ont l'habitude de traiter. Le deuxième thème de la bioinformatique est le stockage, l'analyse et l'exploitation des données. Sous quelle

forme les stocker, dans combien de bases ? Comment les analyser et les exploiter pour arriver à des modélisations et des simulations correctes qui constituent un autre axe de la bioinformatique ? Grâce à certains outils et aux données enregistrées, on commence en effet à modéliser, à simuler des comportements moléculaires ou de structures comme les membranes. Tout le problème ensuite est d'intégrer ces informations, de les mettre à disposition (par l'accès au web par exemple). La fouille de données est enfin une problématique intéressante : de nouvelles connaissances peuvent être déduites à partir des informations stockées et de multiples méthodologies travaillent à cela : mathématiques, statistiques, inférentielles, logiques, voire d'intelligence artificielle. Des recherches sont menées sur de nouveaux algorithmes de fouille et l'intérêt de nombre d'industriels témoigne d'un potentiel économique très important autour de la bioinformatique, même si son développement n'est qu'émergent en France.

Avant de transmettre la parole au premier intervenant, nous pouvons nous arrêter plus particulièrement sur 3 problèmes dont le premier consiste à savoir comment acquérir, stocker et analyser les données des micropuces. De nombreuses approches existent pour comparer deux séries de données. Au sein de notre laboratoire, nous nous intéressons à l'une d'entre elles : celle du modèle de Markov caché qui s'attache à déterminer la probabilité de succession d'une séquence donnée (voir figures 6 et 7).

Les modèles de Markov cachés et microarrays

- Une lame est scannée et stockée dans un fichier sous forme d'une grille de spots dont les informations seront stockées en base de données : l'intensité lumineuse, l'écart type de l'intensité, le bruit, l'écart type du bruit et le nombre de pixels (*i.e.* la taille des spots) →
- Travaux à réaliser : traitement automatisé des images, analyser une lame, comparer les lames entre elles, etc. →
- Moyens : modèles de Markov cachés (HMM) et automates probabilistes. →

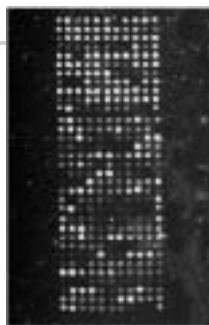


Figure 6

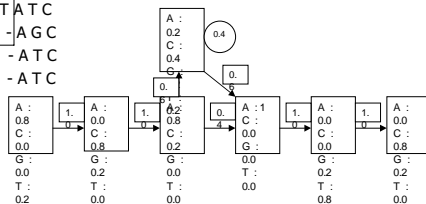
16

Un exemple de modèle de Markov caché en bioinformatique

$S=\{1,2,\dots,q\}$ ensemble d'état, $A=\{A,C,G,T\}$ ensemble d'alphabet de 4 acides nucléiques

Un modèle de Markov cachés est construit à partir de 5 séquences observées

- A C A - - A T G
- T C A A C T A T C
- A C A C - - A G C
- A G A - - A T C
- A C C G - - A T C



$$P(\text{AGTAGC}/\text{Modèle}) = 0.8 \times 1.0 \times 0.2 \times 1.0 \times 0.2 \times 0.6 \times 0.2 \times 0.6 \times 1.0 \times 0.1 \times 0.0 \times 0.2 \times 1.0 \times 0.8 = 0.00036$$

84



Figure 7

L'autre point que je souhaite souligner est la difficulté d'associer connaissances médicales et bases de données génétiques. Prenons l'exemple d'HC forum (<http://HCForum.imag.fr>) qui est une banque de données sur les anomalies chromosomiques de structures familiales et qui apporte depuis 1989 une aide dans le diagnostic et pour le conseil génétique en évaluant le risque de malformation à l'aide de modèles statistiques. Cette base propose un arbre généalogique interactif qui permet aux familles d'être suivies et de connaître les risques génétiques qu'elles encourent en fonction de leurs antécédents. Cette base constitue un véritable pont entre les laboratoires de diagnostic et les centres de recherche du génome.

Enfin, la question reste posée de savoir comment connecter entre elles les bases de données médicales et les bases de données génomiques et protéomiques. Le CHR de Lille dispose d'une base d'1,5 million de dossiers médicaux qui comportent au minimum une cinquantaine d'éléments intéressants de diagnostics ou de résultats de biologie. Comment va-t-on lier ces données médicales qui permettent d'identifier très rapidement des maladies (cancers, diabète, maladies métaboliques, obésité), sur des séries très importantes de patients, aux données de la génomique, de la protéomique, etc. ?

Voici autant de questions auxquelles nous allons tenter d'apporter des éléments de réponse durant cet après-midi de travail.

Dr. Christophe Roumier

c-roumier@chru-lille.fr

LABORATOIRE D'HÉMATOLOGIE DU CHRU

Hôpital Roger Salengro - 59037 Lille Cedex

INSERM UNITÉ 524

Place de Verdun - 59045 Lille Cedex

La plate-forme de génomique fonctionnelle de l'Université de Lille 2 est organisée selon deux grands axes :

- ***l'analyse de l'expression d'ARN*** (transcriptome) par une approche sur puces à oligonucléotides longs.
- ***l'analyse des variations de l'ADN génomique*** par une approche d'hybridation génomique comparative à haute résolution (CGH microarray)

Cette plate-forme bénéficie d'un parc matériel conséquent qui comprend notamment :

- Un spotter Eurogridder qui nous permet de réaliser nos lames à façon
- Un robot MWG BIOTECH pour le réarrangement des plaques oligos avant spotting.
- Un robot pipeteur diluteur Hydra Robotics pour le réarrangement de plaques 96 et 384 puits,
- Un microspectrophotomètre Nanodrop pour le dosage d'ADN et ARN sur micro volumes,
- Un bioanalyzer Agilent pour l'analyse et la quantification des fragments ADN et ARN sur Lab-on-Chips
- Une station d'hybridation,
- Différents types de Scanner pour l'analyse des données

À l'origine, cette plate-forme a été labellisée dans le cadre de la Génopole de Lille-Nord Pas de Calais, axée sur les pathologies génétiques multifactorielles et l'innovation thérapeutique.

L'activité de la plateforme inscrit cette dernière ***au cœur des projets de la Cancéropole Nord Ouest*** qui fédère sur le thème « du dépistage des tumeurs à l'innovation thérapeutique » les CH et U d'Amiens, Lille, Caen et Rouen (G4), les Centres de Lutte contre le Cancer de Caen, Lille et Rouen, les universités de Médecine et des Sciences, les 2 fédérations de cancérologie de Lille et de Caen, les unités INSERM, le CNRS et l'hôpital Erasme de Bruxelles.

Au sein du CHU, un effort important est mené pour améliorer le recueil des informations, aussi bien au niveau des patients qu'au niveau des échantillons de cellules tumorales, ***avec la création d'une tumorothèque et d'une base de données patients, dans le cadre de la cancéropole.***

Ainsi, *au niveau des patients*, le recueil des annotations et de la réponse clinique aux traitements implique une codification permettant leur exploitation informatique. Ce travail est facilité si les patients sont inclus dans des protocoles de recherche médicale où le recueil des informations cliniques est parfaitement standardisé. De même, l'étude approfondie des cellules tumorales génère une quantité astronomique de données : imagerie, morphologie, immunophénotype, cytogénétique, principales anomalies moléculaires, etc. Ces informations doivent être codifiées et intégrées au système commun de gestion des laboratoires du CHU pour créer une

base de données patients fonctionnelle (voir figure 8).

Au niveau des échantillons, **la qualification physique et scientifique** est nécessaire et repose respectivement sur **la qualité de l'échantillonnage** et sur **la qualité des annotations liées au prélèvement**.

La qualification physique est effectuée dans le cadre de la tumorothèque de Lille suivant les recommandations de l'ANAES :

- **Le délai entre la dévascularisation d'une pièce anatomique et sa congélation** est de moins d'une heure et la **traçabilité** de la conservation est complète, avec une **anonymisation** totale de tous les prélèvements.
- Le diagnostic effectué sur les prélèvements permet de qualifier ces derniers en terme de **composition** (pourcentage de cellules tumorales, présence de nécrose ou de tissus sains). Le diagnostic cytologique et/ou histologique permet quant à lui de connaître **la nature des cellules** qui composent le prélèvement.
- Chaque prélèvement (pièces anatomiques, ponctions veineuses, ponctions médullaires...) va générer des produits **secondaires** (ADN, ARN...) par des étapes de tri cellulaire par exemple, aussi bien en cytométrie de flux qu'en microdissection (techniques mises en place à l'Institut de Médecine Prédictive – IMPRT). Cela permet de travailler sur des échantillons de cellules homogènes.

L'annotation biologique et clinique est, quant à elle, la plus complète possible (**qualification scientifique**) : nature et moment du prélèvement, localisation métastatique, rechute, suivi de maladie résiduelle...

L'ensemble de ces informations doit être intégré dans la base de données de la tumorothèque ce qui nécessite **un logiciel puissant d'acquisition et de gestion d'information**, permettant la mise en réseau à la fois **avec les autres cancéropoles**, pour travailler, par exemple, sur les pathologies à prévalence faible, et **avec les bases de données patients**, pour assurer un retour vers le patient du bénéfice des études menées (voir figure 9).

Au sein de la plate-forme de génomique fonctionnelle, la communication entre les différents projets menés est un facteur de progrès majeur. Ces projets concernent différentes investigations en cancérologie :

- les hémopathies malignes, en particulier les leucémies aiguës (lymphoblastiques et myéloblastiques) et le syndrome lymphoprolifératif (LLC et LMNH).
- Les cancers endocriniens et colorectaux familiaux ainsi que les oligodendrogliomes, avec l'étude de la valeur diagnostic et pronostic des délétions qui touchent certains loci.
- Les cancers ORL et dermatologiques, avec la recherche de régions potentiellement oncogènes ou suppresseurs de tumeurs, de profils d'évolution prédictifs d'évolutivité et de dissémination métastatique (voir figure 10).

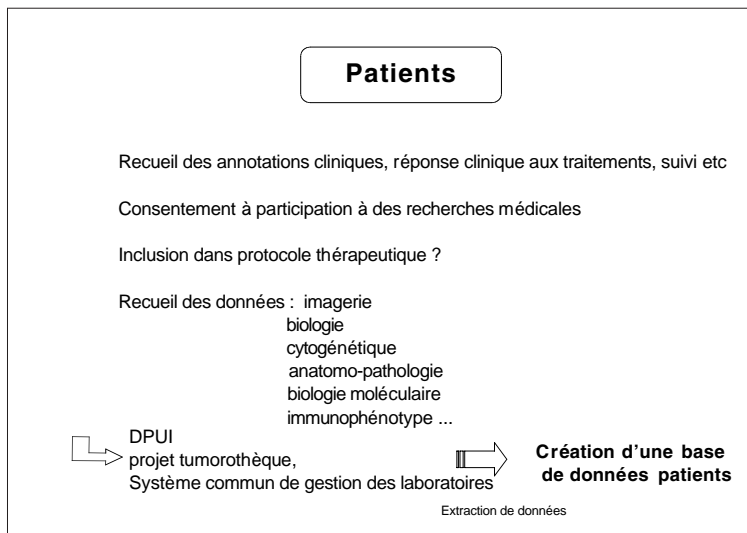


Figure 8

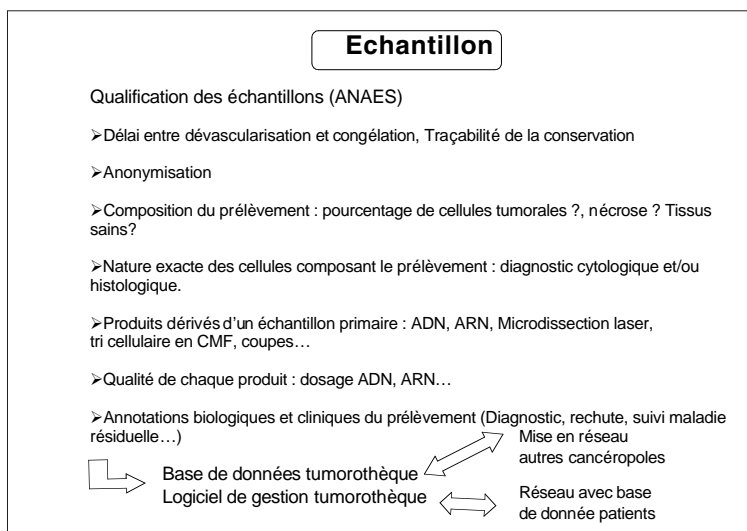


Figure 9

"Analyse des altérations structurales du génome dans les cancers par
hybridation génomique comparative sur puces à ADN"
" Pathologies Multifactorielles et Innovation Thérapeutique"
projet genopole, conseil regional

- Les Hémopathies malignes (C. Preudhomme; C. Roumier; A. Daudignon; JL Lai; JP Kerckaert et al.)
Leucémie aiguë (LAL, LAM) et syndrome lymphoprolifératif (LLC et LMNH)

Unité 524 INSERM Génétique Moléculaire et Approches Thérapeutiques des Hémopathies Malignes, Service des maladies du sang Pr
Bauters, Laboratoire d'hématologie A Pr Cosson, CHRU Lille

- Cancers endocriniens et colorectaux familiaux (Porchet et al.); Oligodendrogliome (Dubois et al)
Valeur diagnostique et prédictive des délétions dans les loci MEN1, MEN2, HNPCC, APC, VHL.
Valeur diagnostique et prédictive des délétions dans les loci 1p et 19q

Unité 377 INSERM Biologie et physiopathologie des cellules mucipares et Laboratoire de Biochimie et de Biologie Moléculaire de l'Hôpital
C.Huriez

- Cancers ORL et dermatologiques (Danzé et coll):

**Recherche de régions potentiellement oncogènes ou suppresseurs de tumeurs, de profils prédictifs
d'évolutivité et de dissémination métastatique.**

Unité 459 INSERM signaux, récepteurs et différenciation cellulaire

Figure 10



PROSKELIA

102, route de Noisy - 93230 Romainville

Tél. 01 49 42 46 67

teresa.garcia@proskelia.com

www.proskelia.com

Teresa Garcia

Responsable de l'identification des cibles, société Proskelia

Proskelia a été créée sur le modèle d'une « mini-pharma » en intégrant toutes les fonctions nécessaires à la découverte de nouveaux médicaments, de la recherche exploratoire jusqu'aux premières phases du développement clinique. Elle se compose aujourd'hui d'une centaine de personnes (voir figure 11).

Notre entreprise est le fruit de l'externalisation du domaine des maladies des os d'Aventis Pharma. Nous menons nos travaux en collaboration avec un réseau académique, des prestataires de services, et d'autres entreprises pharmaceutiques. La chaîne de valeur qui mène à la découverte d'un médicament part de l'identification et de la validation de cibles. Il s'agit ensuite d'effectuer un criblage à haut débit pour identifier des molécules d'intérêt dont il reste à vérifier les fonctions biologiques, *in vitro et in vivo* (voir figure 12).

23

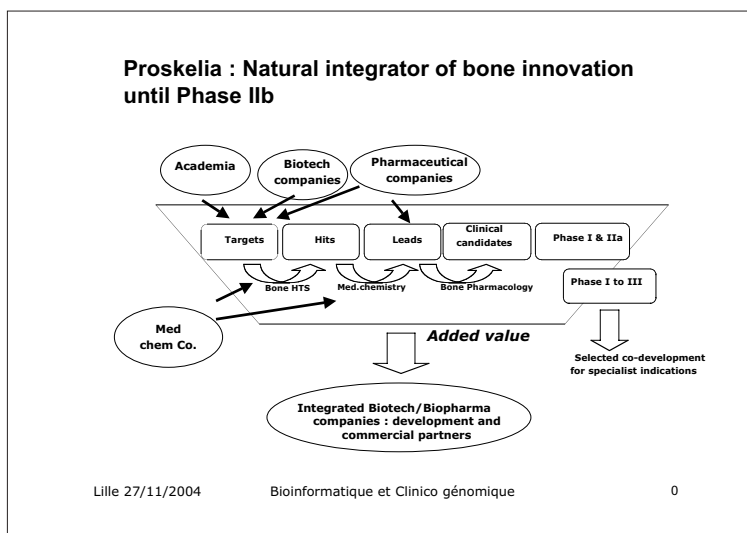


Figure 11

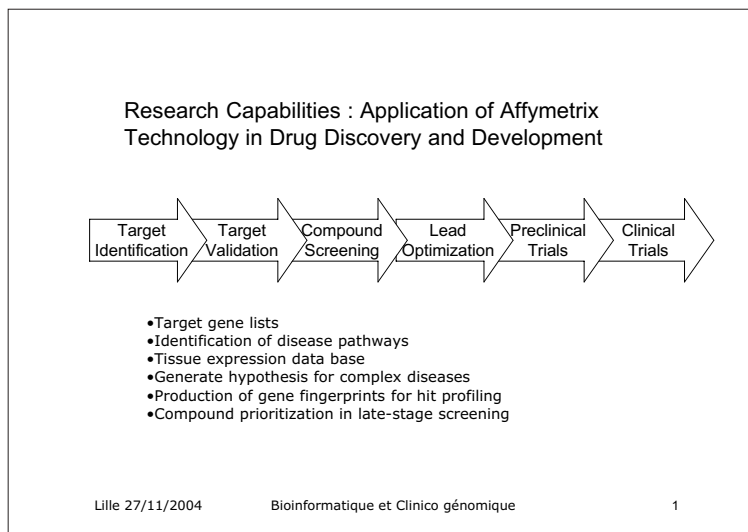


Figure 12

24 **L'identification de cibles** a pleinement recours à la génomique et en particulier aux puces ARN pour la compréhension des voies métaboliques et des relations entre protéines qui conditionnent les réponses biologiques. L'apport de la bioinformatique est à ce stade important pour cribler rapidement l'expression d'une cible à travers un grand nombre de tissus.

En aval, ces outils vérifient que la réponse induite dans un type cellulaire par des molécules chimiques correspond bien aux objectifs recherchés par rapport à un contrôle positif connu. Cela permet de donner la priorité à certaines molécules en terme de développement pharmaceutique.

Au sein d'Aventis, nous avons mené une expérience comprenant un grand nombre d'échantillons, sur des modèles d'ostéof ormation de différentes espèces. Des collaborations extérieures pour la collecte

d'échantillons et plusieurs modèles *in vitro* ont été mis en place (voir figures 13 et 14).

Les échantillons ont été analysés par la technologie Affymetrix, qui présentait alors certains inconvénients. Le génome de la souris (40 000 gènes) était en effet réparti à l'époque sur 5 chips et la confiance que l'on accordait aux résultats obtenus était modérée. Cela nous obligeait à dupliquer les hybridations. Pour chaque point expérimental, 10 hybridations et autant de mesures étaient donc nécessaires : l'analyse s'avérait donc relativement lourde. Aujourd'hui, les génomes de la souris et de l'homme sont concentrés sur une puce par génome. La confiance que l'on accorde aux mesures est dorénavant plus importante car les sondes sont de meilleure qualité. À l'époque, nous avons été contraints de développer un outil en interne pour gérer les données brutes que nous obtenions,

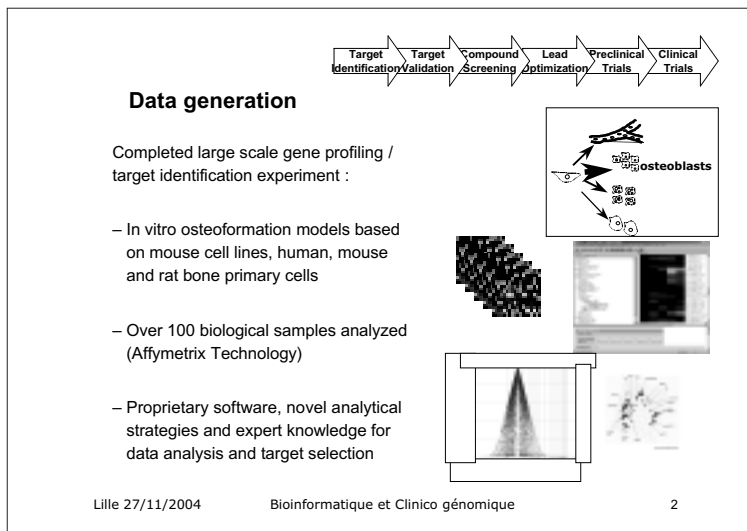


Figure 13

25

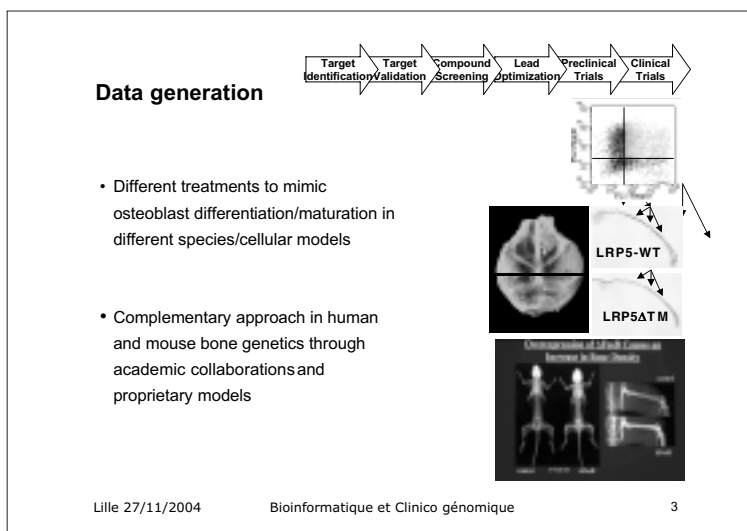


Figure 14

pour pouvoir les annoter et les analyser statistiquement. Aucun outil sur le marché ne permettait alors de calculer de simples moyennes de valeurs mesurées sur deux puces.

Grâce à cet ensemble de technologies, nous avons pu identifier des gènes régulés de manière statistiquement significative. Au sein de ceux-ci, nous avons ensuite sélectionné des cibles potentiellement intéressantes en terme de développement de têtes de série chimiques et nous les avons utilisées pour effectuer des tests de criblage à haut débit. Cette méthodologie nous oblige à multiplier les expériences et à sélectionner très strictement les cibles rencontrées (voir figure 15).

L'un des défis qu'a du relever Proskelia au moment de sa création était d'adapter à son échelle l'ensemble des moyens hérités d'Aventis dans le domaine de la gène-

mique. Nous avons réussi à passer d'une équipe complète de bioinformaticiens à une seule personne bénéficiant d'un appui technique et des logiciels adéquats. Cela a notamment été possible grâce à la qualité des puces qui s'est grandement accrue et aux bases de données du domaine public désormais beaucoup plus fiables et régulièrement mises à jour. La fédération des données est ainsi devenue un exercice plus rapide et mieux maîtrisé (voir figure 16).

Notre base de données, Bone Biology Database, nous donne accès à un large éventail d'informations ainsi qu'aux bases de données publiques et aux données analysées par le logiciel Geko (Gene Expression Knowledge Organization) qui permet de rassembler, normaliser et filtrer les données ainsi que d'effectuer des calculs statistiques et de visualiser nos résultats (voir figure 17).

26

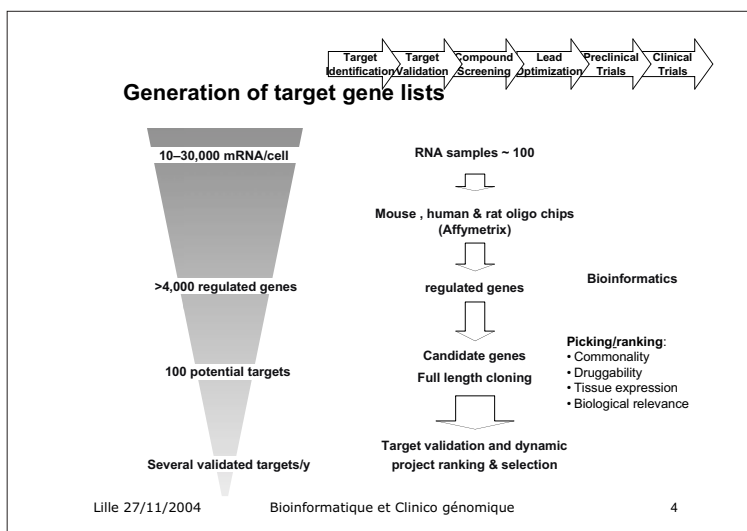


Figure 15

Data analysis solutions : annotations

CGC Aventis Team - 10 FTEs

- Cluster all sequences – GB, Incyte, CGC clone libraries
- Make searchable database of consensus sequences
- BLAST AFFX sequences against this internal consensus DB
- Costly – clustering software, time for clustering, time for BLAST – new machine 10 processors.
- Clustering software unreliable (CAT tools from Pangea) – clusters similar (but different) genes together. E.g. from different species

PSK - 1 FTE + IT support

- Use Public Domain EST clustering program – UniGene
- Direct lookup from qualifier to UniGene cluster ID
- From UG ID, have links to other gene data sources
- Enables cross-species data retrieval through Gene Symbol
 - not possible before

Lille 27/11/2004

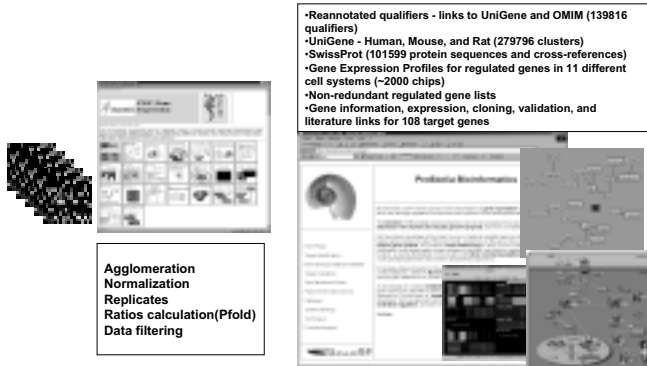
Bioinformatique et Clinico génomique

5

Figure 16

27

Data analysis solutions : user interface



•Reannotated qualifiers - links to UniGene and OMIM (139816 qualifiers)
 •UniGene - Human, Mouse, and Rat (279796 clusters)
 •SwissProt (101599 protein sequences and cross-references)
 •Gene Expression Profiles for regulated genes in 11 different cell systems (~2000 chips)
 •Non-redundant regulated gene lists
 •Gene information, expression, cloning, validation, and literature links for 108 target genes

Agglomeration
 Normalization
 Replicates
 Ratios calculation(Pfold)
 Data filtering

Lille 27/11/2004

Bioinformatique et Clinico génomique

6

Figure 17

Au-delà de la découverte de cibles, la génomique peut aussi aider à la sélection de nouvelles molécules thérapeutiques. Le profil biologique de molécules issues des tests secondaires de criblage est établi sur un échantillonnage de gènes représentatifs, sélectionnés à partir de nos données génomiques. La génomique nous aide ainsi à mieux sélectionner et à donner une

nouvelle valeur ajoutée à des molécules chimiques (voir figure 18).

L'important est ainsi de pouvoir fédérer des ensembles de données, de pouvoir exploiter rapidement des expériences très hétérogènes dans leurs formes et leurs contenus (voir figure 19).

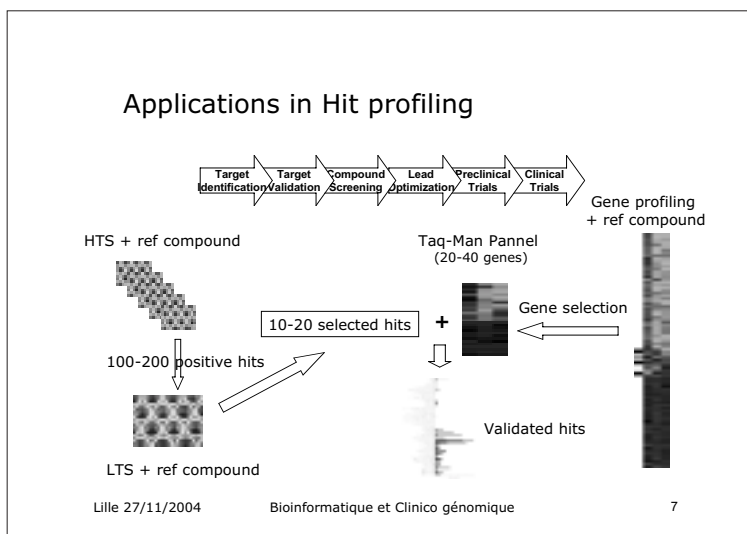
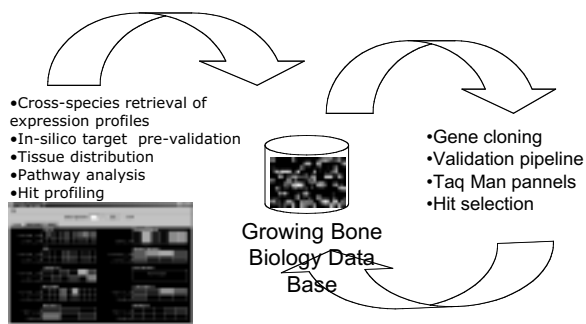


Figure 18

A dynamic target/hit selection through gene profiling



Lille 27/11/2004

Bioinformatique et Clinico génomique

8

Figure 19



IBM DIVISION SCIENCES DU VIVANT
Tour Descartes - La Défense 5
2, av. Gambetta - 92066 Paris La Défense Cedex 50
Tél. 01 49 05 65 24
psempe@fr.ibm.com

Pascal Sempé

**Business Development Executive, IBM
Division Sciences du Vivant**

La médecine de demain consistera à intégrer et fédérer des données cliniques, avec des données génomiques et démographiques, pour générer de l'information, par la statistique et l'épidémiologie (voir figure 20).

Au cours de la vie d'un patient, un certain nombre d'informations va être généré : terrain génétique, observations cliniques, données familiales, etc. Nombre d'acteurs

vont avoir un intérêt à accéder à ces données : les cliniciens, le secteur hospitalier, les biobanques, l'industrie pharmaceutique, les laboratoires d'analyse, les cabinets de radiologie, etc. Grâce à l'ensemble de ces informations, l'objectif est d'aboutir à une **médecine personnalisée**, où il est possible d'évaluer la réponse à certains traitements voire de prédire l'évolution de l'état de santé d'une personne. Une personne identifiée avec un facteur de susceptibilité vis-à-vis d'une pathologie pourra ainsi être suivie en conséquence (voir figure 21).

31

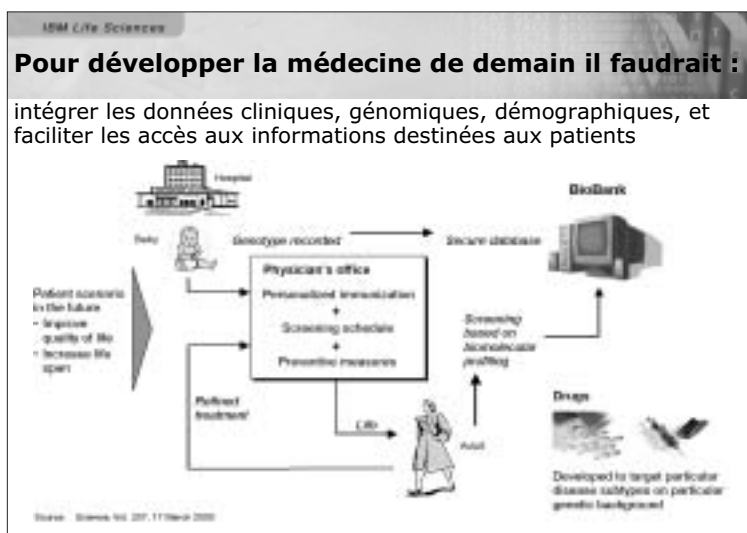


Figure 20

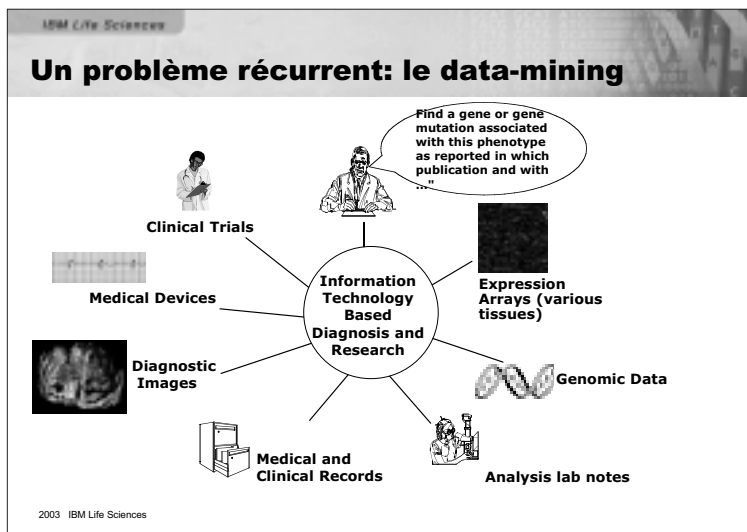


Figure 21

32 Ces nombreuses données peuvent rapidement submerger les professionnels, notamment lorsqu'elles n'ont pas été organisées et fédérées de façon homogène. Deux solutions existent :

- Intégrer l'ensemble des informations dans une base de données structurée. Il faut alors définir la structure de la base de données et résoudre des problèmes de sémantique.
- Gérer les données telles qu'elles sont et trouver des solutions d'interconnexion qui permettent de poser des requêtes croisées

Aujourd'hui, il est possible d'accéder à **un grand nombre de bases de données qui ne sont cependant pas interconnectées**. Il s'agit donc à chaque fois de connaître leur langage propre, leur structure et les moyens d'interrogation disponibles. L'information, si elle existe, reste donc très difficile à exploiter (voir figure 22).

La fédération de ces bases doit résoudre le problème de leur hétérogénéité et permettre à chaque métier de gérer sa base de données tel qu'il l'entend. La masse d'informations doit apparaître de façon **transparente**, comme si elle était stockée en un endroit unique (voir figure 23).

Une fois ces données fédérées et rendues disponibles, se pose le problème de l'infrastructure. La tendance va actuellement vers des systèmes où les bases de données virtuelles et des applications standardisées deviennent accessibles à la plupart des acteurs de l'environnement du patient. Dans le milieu hospitalier, les volontés s'affirment de regrouper les infrastructures informatiques pour homogénéiser les bases de données entre services médicaux.

L'évolution de l'informatique est nécessaire, pour répondre aux pics de demande

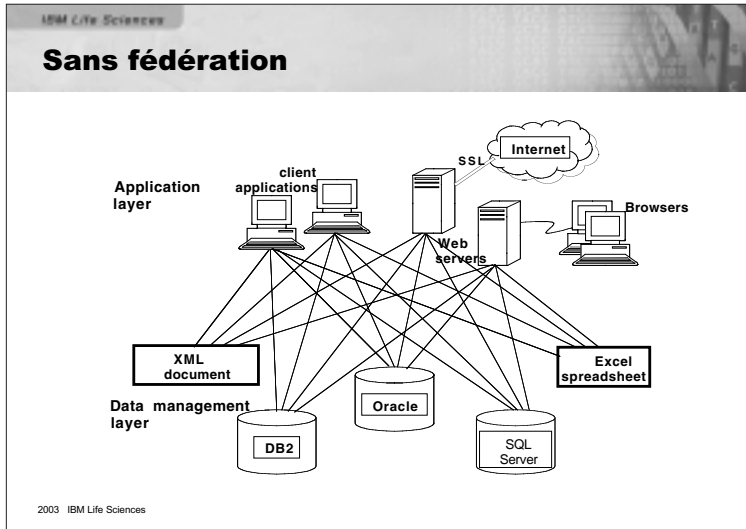


Figure 22

33

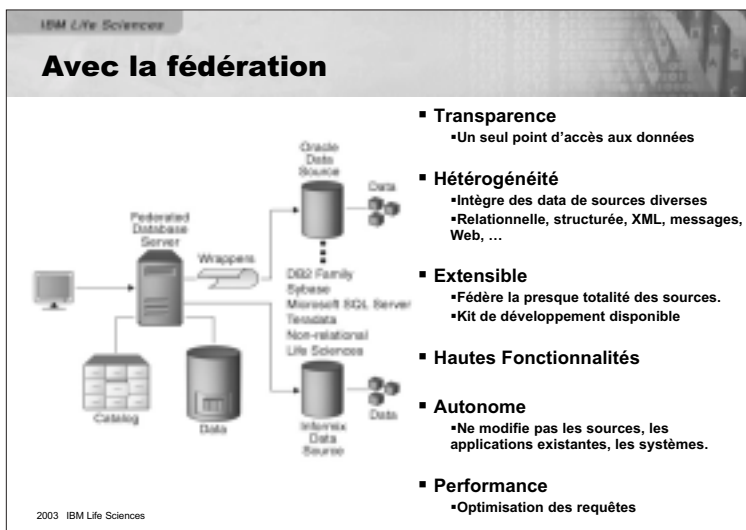


Figure 23

de calcul ou de disponibilité d'applications. Le système idéal serait un système qui s'adapte à la demande, où la mutualisation des ressources s'adjoint à la mutualisation des données, ce qui correspond au **concept « on demand » d'IBM** (voir figure 24).

Dans ce système, l'infrastructure informatique mise à disposition d'une communauté s'adapte à la demande de calcul. Le système va chercher les ressources là où elles sont disponibles et répartir la puissance de calcul en fonction de la demande. L'informatique devient alors une facilité, comme l'eau où l'électricité, sur la base **d'un coût cohérent avec la consommation**.

À la base de ce système, le **grid computing** (grille de calcul) tisse, via le réseau, des rela-

tions entre les centres de calcul et les fournisseurs d'applicatifs. L'observation des systèmes d'information des entreprises ou hôpitaux montre que de nombreuses machines constituent un vivier important de puissance de calcul et de stockage.

Physiquement, vous allez chercher la puissance de calcul, là où elle est disponible, lorsque vous en avez besoin. Le calcul peut être distribué sur plusieurs sites, comme les données. L'infrastructure informatique peut être dématérialisée (voir figure 25). Le programme innovant **« e-diamond »**, centré sur le dépistage du cancer du sein en Grande-Bretagne, réunit IBM, Mirada, un éditeur de logiciel, des universités et des hôpitaux (voir figure 26).

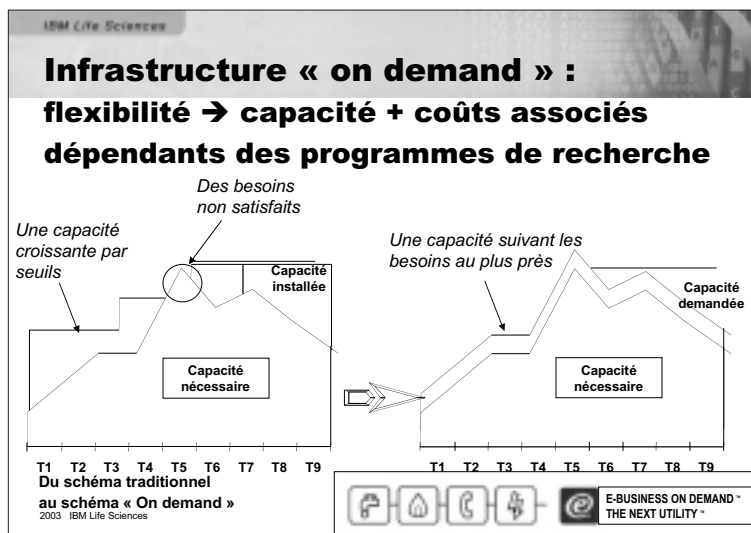


Figure 24

IBM Life Sciences

Server/Storage Utilization

	Peak-hour Utilization	Prime-shift Utilization	24-hour Period Utilization
Mainframes	85-100%	70%	60%
UNIX	50-70%	10-15%	<10%
Intel-based	30%	5-10%	2-5%
Storage	N/A	N/A	52%

Source: IBM Scorpion White Paper: Simplifying the Corporate IT Infrastructure, 2000

2003 IBM Life Sciences

Figure 25

35

IBM Life Sciences

Un programme innovant en Grande-Bretagne: e-Diamond

MIRADA™ SOLUTIONS

UCL

Oxford Radcliffe Hospitals NHS Trust

KING'S College LONDON

St George's

IBM

Scottish Breast Screening Programme

2003 IBM Life Sciences

Figure 26

Dans ce programme qui a débuté en 1997, 1,5 million de femmes entre 50 et 60 ans sont dépistées tous les 3 ans, sur 92 centres de dépistage, soit environ 230 radiologistes (voir figure 27).

Ce programme bénéficie d'un environnement favorable, caractérisé par :

- L'engagement du gouvernement britannique dans l'informatisation des dossiers médicaux.
- L'amélioration de la diffusion de la culture informatique au sein des équipes médicales
- La qualité des techniques de normalisation et de capture des images par les Picture Acquisition and Communication Systems (PACS)

Le Breast Screening Program (BSP) a ainsi permis de dépister plus de 11,000,000 femmes et détecté plus de 65,000 cancers, sauvant 230 vies/an. Son objectif d'ici à 2010, sauver 1250 vies/an, se heurte à la pénurie de spécialistes dans certaines régions.

Les cliniciens sont submergés de données et ne sont pas formés à gérer un dossier informatique de cette dimension. Il faut mettre en place une infrastructure et des outils qui puissent extraire les informations pertinentes pour un diagnostic sûr et rapide.

Chaque cliché génère en effet 3 millions d'images par an, de taille importante (entre 25 et 40 Mbytes) dont la définition est cruciale pour l'interprétation. La compression doit se faire sans perte. Ces mammographies sont envoyées sur une grille de calcul avec les données « patients » associées. Pour réaliser des comparaisons, chaque image est normali-

sée. En effet, le cliché peut avoir été pris sous des angles différents, par des appareils de différents constructeurs.

Le SMF fournit aux cliniciens des outils pour l'aider au diagnostic (ESAD). Ces programmes vont analyser les images au fur et à mesure qu'elles sont produites et avertir le clinicien en cas de signal anormal. Pour ce faire, on construit par exemple des visualisations 3D de très grande qualité (voir figure 28).

Les grilles de calcul recèlent de potentialités importantes : les images collectées tous les 3 ans sur une même patiente vont pouvoir aider à définir des schémas types d'évolution des tumeurs et des nodules. Cela pourra permettre, par exemple, d'adapter le traitement en fonction du schéma-type observé. Ces informations peuvent être aussi mises en relation avec d'autres données collectées : biologiques, génétiques, etc.

IBM Life Sciences

Le Rationnel pour le dépistage

- Diagnostique rapide = meilleur pronostic
- Détection à 0.5cm a un pronostic favorable dans 99% des cas; mais dès >2cm tombe à 50%
- Quand l'observation est faite par la patiente, la tumeur approche en général le cm, soit en général après 6-8 ans de croissance
- 97% de chances de survie au delà de 5ans pour une tumeur localisée. Pour une tumeur métastasée: 77%

2003 IBM Life Sciences

Figure 27

37

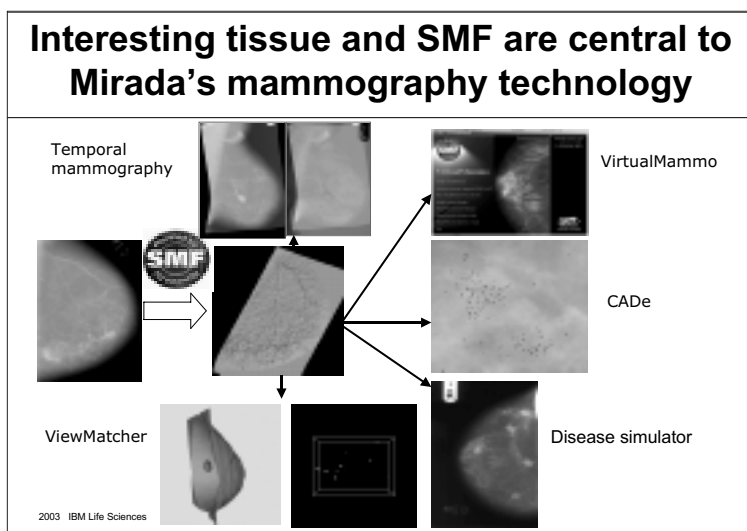


Figure 28



IBM DIVISION SCIENCES DU VIVANT

Tour Descartes - La Défense 5

2, av. Gambetta - 92066 Paris La Défense Cedex 50

Tél. 01 49 05 65 24

thierry_peuchot@fr.ibm.com

Thierry Peuchot

***Business Development Executive, IBM
Division Sciences du Vivant***

L'initiative d'IBM dans les Sciences de
la Vie

« Vers une médecine personnalisée »

La volonté d'IBM de se positionner sur le marché des Sciences du Vivant s'est traduite par une initiative stratégique et mondiale. Cette initiative partie des Etats-Unis remonte à trois ans. IBM avait été sollicité par un acteur majeur du Projet de Séquençage Humain, the Human Genome Project. Au début la demande portait sur des besoins en calculs intensifs, et a rapidement évolué vers du stockage, de la gestion et l'interprétation des données, de la gestion de workflow et bien sûr vers des niveaux de sécurité élevés. L'équipe constituée pour aider à cette convergence du monde biologique et du monde informatique est le point de départ de l'équipe IBM Santé Sciences du Vivant qui ne cesse de s'accroître.

La division d'IBM Santé Sciences du Vivant est une entité se focalisant sur la mise en place de solutions destinées à aider les chercheurs confrontés à l'accumulation de données qu'ils se doivent de gérer, de corréler, d'interpréter...

Cette division adresse les besoins des segments de marché suivants :

- Les laboratoires pharmaceutiques ;
- Les compagnies de biotechnologies,

spécialisées en génomique et protéomique ;

- Le secteur académique ;
- Les acteurs du diagnostic médical ;
- Les acteurs de l'instrumentation médicale et
- La santé

Aujourd'hui IBM est encore souvent perçu comme un constructeur fournisseur de technologies matérielles. En fait, plus de 50 % de son activité est réalisé dans les services au travers notamment de sa division Consulting (Business Consulting Services). À côté de cette activité dans l'univers des services, IBM bien sûr continue de jouer un rôle important dans la fourniture de plates-formes informatiques et de logiciels d'infrastructure.

IBM Santé Sciences du Vivant en tant qu'entité « Solutions » s'attache à concevoir des solutions répondant aux besoins fonctionnels de ses clients. Une solution type comprend ainsi une partie « services », du matériel, une partie « logiciels d'infrastructure » ainsi qu'une ou plusieurs « briques logicielles » venant d'éditeurs tiers. Une décision stratégique a conduit IBM à se retirer des applicatifs verticaux ou applicatifs métiers. Le catalogue logiciel d'IBM se limite aux logiciels intermédiaires ou « d'infrastructure » facilitant la communication entre applications, la gestion des bases de données et l'administration des plates-formes.

Dès lors, on comprend le rôle important des partenariats avec les éditeurs dans la conception de ces solutions. Nos équipes interviennent en tant que « chef d'orchestre » pour la mise en place de solutions : services, consulting, intégration technique, strategic outsourcing...

Pour nous aider dans la conception de ces solutions innovantes nous sollicitons également un des huit laboratoires de recherche dont IBM dispose de par le monde. Chacun de ces huit sites dispose de domaines d'expertises propres. Le site de Watson, par exemple, dispose d'un centre de bioinformatique au sein duquel travaillent plus de 70 Ph. D. avec de solides expériences dans le domaine de la R&D.

C'est notamment au sein de ce laboratoire qu'est conduit le projet Blue Gene qui vise à construire une nouvelle famille de supercalculateurs optimisés en termes de bande passante, de « scalability » et de capacité à traiter une quantité énorme de données. Tout cela en ne demandant qu'une partie de l'énergie et de l'espace des supercalculateurs actuels les plus rapides. Parmi les premières applications qu'IBM compte maîtriser avec ce nouveau système figure le pliage Ab Initio des protéines humaines, avec l'espoir d'offrir aux chercheurs une meilleure compréhension des maladies et de permettre de trouver un traitement approprié. En Novembre 2003, un premier prototype de Blue Gene a été présenté. De la taille d'un téléviseur de 80 cm environ ce prototype dispose d'une puissance de 2 teraflop, ce qui le place au 73^{ème} rang du classement TOP 500 des supercalculateurs.

Concernant le monde médical, l'évolution des connaissances oblige à interpréter de plus en plus de données. Dans les années

30, un médecin établissait son diagnostic à partir de données sommaires sans investigation complémentaire. Nous disposons aujourd'hui d'informations sur les organes, sur les tissus, au niveau de la cellule, voire au niveau moléculaire. Pour réunir et aider à interpréter l'ensemble de ces données, l'informatique devient indispensable.

Le concept « **Information Based Medicine** » correspond au support apporté aux médecins dans l'établissement de leurs diagnostics à partir de données numériques complémentaires : imagerie, résultats biologiques, etc.

L'« Information Based Medicine » est le processus destiné à améliorer les pratiques pharmaceutiques et médicales grâce à la connaissance résultant de l'intégration de données cliniques et biomédicales. L'« Information Based Medicine » va permettre aux acteurs de la Santé (Patients, Professionnels de santé, Payeurs et les Institutions qui les représentent) d'améliorer le diagnostic et le traitement des maladies en facilitant/accélérant la mise en place de thérapies ciblées et donc de soins personnalisés tels que décrits dans l'étude réalisée par notre division Consulting BCS « Pharma 2010 ».

Sur les 5 dernières années, l'informatisation a grandement progressé. Les différents services de l'hôpital sont capables d'échanger point à point des informations, ce qui évite les transports physiques de dossiers. Avec l'arrivée d'Internet, ce modèle a encore évolué. Nous passons d'un monde hospitalo-centrique à un espace dématérialisé où l'important, pour l'utilisateur, est d'accéder à l'information dont il a besoin. L'hôpital a un rôle majeur à jouer et doit être capable d'échanger des

données avec les praticiens, les assurances, voire d'autres services tels que les services de veille sanitaire, de transports et d'ambulances...

Aujourd'hui, les systèmes d'information hospitaliers sont encore souvent organisés en silos ce qui conduit à une allocation des ressources non optimale : une majorité d'acteurs « colmate les brèches » et assure les mises à jour quotidiennes de leur système où chaque modification doit être répercutée de façon répétitive. En terme d'exploitation, la gestion de ces processus cycliques est gourmande en ressources. Notre objectif dans ce domaine est d'aider à réorganiser les systèmes d'information pour mutualiser les ressources, mieux les partager et centraliser certaines fonctions, voire mettre en place des systèmes « à la Demande » capables de suivre au mieux l'évolution des besoins de nos clients...

Actuellement, un autre point clé est la fédération et l'accès à l'ensemble des données disséminées et hétérogènes. Pour ce faire, nous disposons d'un logiciel d'infrastructure, appelé **DiscoverLink**, qui fédère et mutualise des sources de données très disparates, de par leur contenu, leur localisation et leur format (textes, images, fichiers Excel, bases de données etc...). **DiscoveryLink** offre à l'utilisateur un accès unique à l'ensemble des données, comme si les sources d'informations étaient regroupées au sein d'une base de données **virtuelle et unique**. L'interrogation des bases ainsi fédérées s'en trouve facilitée et accélérée : grâce à une requête qui adresse simultanément l'ensemble des bases. Une synthèse des résultats de cette requête est ensuite présentée à l'utilisateur.

Nos chercheurs sont souvent sollicités sur des thématiques très pointues. Citons deux exemples : L'université UCSF (L'University of California San Francisco) et l'hôpital Fred Hutchinson.

L'UCSF souhaite accélérer le développement des traitements personnalisés ou pharmacogénomiques en intégrant/fédérant des données de recherche clinique issues de sources multiples. Pour ce faire, un système de « Clinical Genomic Information Mining » est en cours de mise en place afin de faciliter l'interrogation, la fouille et l'analyse simultanées des données générées par les chercheurs et les données cliniques de l'hôpital. Ce système doit permettre la formulation de nouvelles hypothèses quant aux marqueurs génétiques et biologiques des maladies et donc la mise en place de diagnostics plus précis et des traitements personnalisés, tout ceci pour le plus grand bénéfice du patient.

L'hôpital Fred Hutchinson souhaitait réorganiser son système d'information, pour pouvoir répondre aux questions suivantes :

- Pourquoi certains sujets répondent-ils mieux aux vaccins ?
- Pourquoi certains sujets sont-ils plus sensibles aux cancers ?

Au sein du programme Shared University Research, IBM a mis à disposition de cet hôpital des experts, du matériel et des logiciels et a contribué au financement du projet au cœur duquel se trouve Discoverylink. La contribution financière d'IBM peut ainsi être utile pour amorcer certains projets et est attribuée en fonction des critères suivants :

- Programme commun à plusieurs universités
- Caractère innovant de la recherche.

En résumé

L'investissement stratégique d'IBM dans le domaine de la Santé et des Sciences du Vivant porte sur le développement de pratiques de conseil, la conception de solutions verticales personnalisées visant à transformer les hôpitaux et les industries connexes, la proposition de nouvelles méthodes pour accélérer la recherche médicale et optimiser les processus, tout en diminuant le coût des soins et de R&D.

Chaque nouvelle solution capitalise sur la connaissance du secteur et sur les méthodologies de conseil d'IBM Business Consulting Services, combinées avec les compétences et ressources technologiques de la Division Santé Sciences du vivant d'IBM.

Discussion

Régis Beuscart

Quel est le problème de stockage ou de gestion de données le plus important que vous ayez été amené à résoudre ? Avez-vous eu recours, pour le calcul, à d'autres ordinateurs que ceux de votre entreprise ?

Teresa Garcia

En 1999, pour retrouver les annotations d'un grand nombre de gènes, nous avons été obligés de faire appel à l'équipe complète d'un laboratoire spécialisé. Aujourd'hui, ces informations sont plus accessibles.

Pascal Sempé

Le problème le plus difficile est de comprendre les différents langages de ces communautés. La définition, la sémantique doivent être acceptées par chacun. Chaque information est produite par des professionnels qui en ont une vision différente.

Régis Beuscart

Nous venons de parler des liaisons entre bases de données médicales, biologiques, hospitalières, etc. Cela ne va pas sans soulever des problèmes de confidentialité, de traçabilité, d'échanges de ces données qui intéressent beaucoup d'intervenants y compris les assureurs. Si je prends l'exemple du CHRU de Lille, on ne peut évidemment pas accéder aux bases de données de psychiatrie, de fertilité, des IVG. Les contraintes sur l'information de nature médicale ou biologique personnalisée sont fortes. Est-ce que cela peut gêner ce type de déploiement, des solutions existent-elles, y compris pour les essais thérapeutiques ?

Teresa Garcia

Je pense en effet qu'il existe désormais des

systèmes de codage de données assez performants.

Pascal Sempé

Ce problème doit effectivement être soulevé et c'est la raison pour laquelle la plupart des données sont anonymisées. Cette anonymisation n'est évidemment pas complète puisqu'il faut pouvoir revenir du patient aux résultats des analyses biologiques par exemple. Mais un certain nombre de solutions de sécurité existent en effet. Nous avons, sur des programmes multicentriques, déployé un certain nombre de stratégies de cryptage et de codification des données permettant de définir ce que l'on met ou non à la disposition de la communauté : les données sensibles, prioritaires, etc. La sécurité des données est cependant un domaine qui évolue très vite et où rien n'est jamais acquis.

Christophe Roumier

Le problème vient du plan expérimental et de la constitution par exemple des tumorothèques. Le consentement du patient devient nécessaire. L'anonymisation résout une partie des problèmes : le retour s'opère vers les services cliniques participant aux études qui doivent prendre en charge cette partie éthique.

Guillaume Lièvre

Quelle est la croissance aujourd'hui des tumorothèques et des tissuthèques ? Comment vont-elles être reliées à la partie clinique, c'est à dire au dossier patient ?

Christophe Roumier

La tumorothèque des lymphopathies malignes du CHR croît de 400 nouveaux diagnostics par an. Le nombre d'échantillons est très variable suivant le type de

pathologies. Dans la plupart des CHU, il existe aussi des banques de cryopréservation de tissus et de cellules mais ces prélèvements ne sont pas forcément fléchés « tumorothèque ». Ils pourront bénéficier d'une qualification a posteriori mais leur mise à disposition dans les tumorothèques présuppose qu'il y ait eu un consentement du patient et que toutes les qualifications ANAES aient été suivies, ce qui est loin d'être le cas. Il faut plutôt considérer un aspect prospectif que rétrospectif.

Régis Beuscart

Au niveau des tumorothèques, des accumulations de tissus ont été réalisées dans des conditions de conservation très différentes. Sur le plan informatique, l'indexation de ces éléments, en relation avec les dossiers patients de l'hôpital est un problème complexe, qui implique trois ou quatre bases différentes (laboratoires, anatomo-pathologie, Centre Oscar Lambret).

L'architecture même d'une tumorothèque n'est pas simple : les informations enregistrées varient en fonction du type d'échantillon analysé. La mise en relation avec les bases données cliniques et médicales peut être épineuse, notamment dans le cas de sites distants. La récupération en rétrospectif pose des problèmes d'organisation physique.

CLINICOGÉNOMIQUE,
VERS DE NOUVELLES
MÉTHODES
BIOINFORMATIQUES
D'INTERPRÉTATION DES
DONNÉES CLINIQUES



GENFIT

Parc Eurasanté

885, avenue Eugène Avinée - 59120 Loos

Tél. 03.20.16.40.00 - Genevieve.Martin@genfit.com

www.genfit.com

Dr Geneviève Martin

Chef de projet et coordinatrice

Nouveaux Développements, Genfit

Créée en 1999, Genfit mène des travaux de recherche en partenariat avec des industriels de la pharmacie et a développé en interne une activité de R&D centrée sur les maladies métaboliques. Basée sur le parc Eurasanté, elle a développé deux filiales :

- CardiAlpha, qui conçoit et gère des essais cliniques
- IT.Omics, société de bioinformatique dont les compétences se sont révélées rapidement indispensables pour l'analyse des données issues de nos microarrays.

Du point de vue du biologiste, le domaine de la bioinformatique apparaît complexe mais apporte des solutions indispensables. Ainsi, les données protéomiques sont différentes des données génomiques : un gène peut être exprimé sans entraîner de différence fonctionnelle au niveau de la protéine. L'ensemble de ces signaux est caractérisé de façon linéaire (signal faible ou fort) et comparé à des données non-linéaires : tension artérielle, développement d'une plaque d'athérosclérose, etc. Une telle comparaison porte donc sur des éléments différents qui n'ont pas les mêmes unités de mesure et qui varient dans le temps. La bioinformatique permet de passer de cet ensemble de données élémentaires non-ordonnées à une organisation stable, construite autour de rela-

tions de nature biologique (voir figure 29).

Les biologistes doivent ensuite pouvoir interpréter ces données. Sur ce point, la visualisation des données reste un chantier important : comment représenter des notions comportant plus de 4 dimensions ? IT.Omics a mis à notre disposition l'outil LSGraph® qui permet d'établir des liens essentiels entre deux gènes à partir des données de la littérature. Pour atteindre notre objectif, à savoir le développement de méthodes de diagnostics et de traitements, il est désormais nécessaire d'aller plus loin en passant des données d'expression aux données cliniques. Après l'identification d'un gène, il s'agit dorénavant de modéliser ses interactions et mécanismes de régulation pour déterminer la molécule la mieux adaptée à agir sur le site actif de la protéine, par exemple et de ce fait, le médicament à développer (voir figure 30).

Les présentations qui suivent portent sur deux aspects très différents :

- ***La corrélation des profils d'expression génique avec les données cliniques des patients*** afin (1) d'identifier les signatures moléculaires des pathologies tels que le cancer et les maladies métaboliques en vue de mettre au point des outils de diagnostic et de pronostic, (2) de découvrir de nouvelles cibles thérapeutiques, de nouvelles classes biologiques et voies métaboliques

afin de développer de nouvelles molécules pour des traitements mieux adaptés.

- **la corrélation et le partage des données de génotypage avec les données cliniques des patients.**

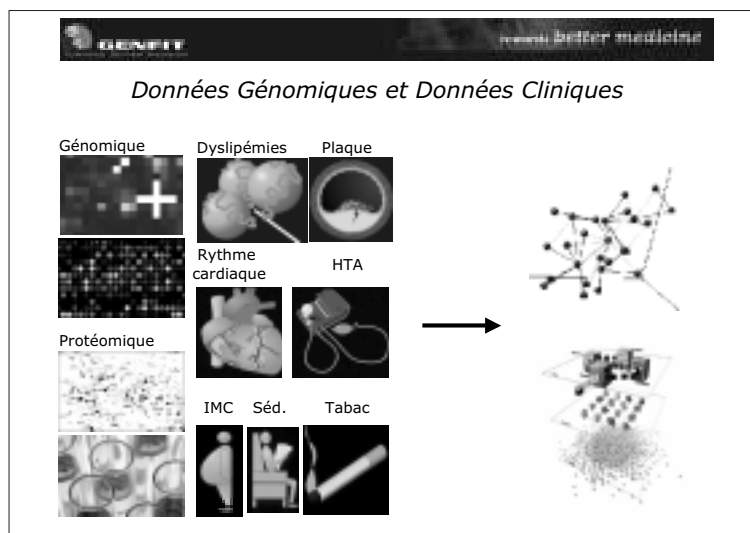


Figure 29

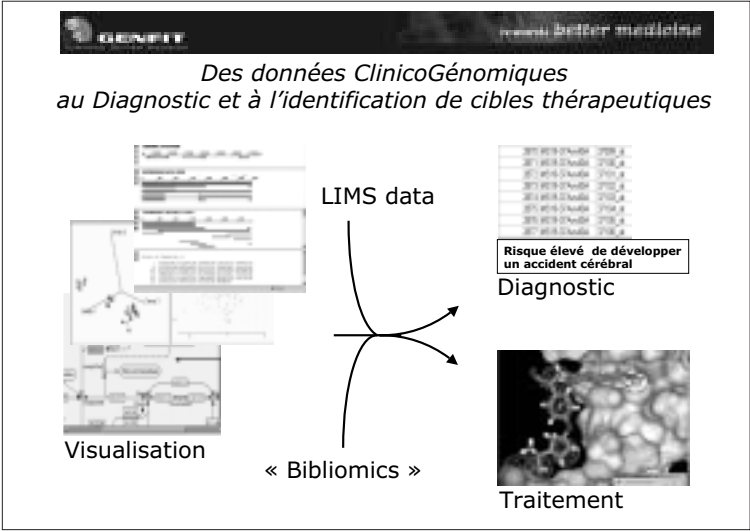


Figure 30



IT.OMICS

Parc Eurasanté

885, Avenue Eugène Avinée - 59120 LOOS

Tél. 03.20.16.40.50 - guillaume.lievre@it-omics.com

www.it-omics.com

Guillaume Lièvre

Directeur Général, IT.Omics

Des données clinico-génomiques à la cible thérapeutique :

Exemple d'intégration d'une plateforme bioinformatique dans la chaîne de recherche.

IT.Omics contribue à la mise au point de nouveaux médicaments, en aidant à travers son expertise et ses solutions à la transformation des données biologiques et cliniques en connaissance des mécanismes pathologiques.

Le processus de mise au point d'un médicament, dans son nouveau paradigme (« target-based drug discovery ») ne vise plus le traitement des symptômes, mais l'identification et le traitement des causes des pathologies. Ce processus reste un processus en deux grandes phases : Recherche, puis Développement Clinique. Cependant la phase de Recherche, qui était autrefois essentiellement une phase d'essai et de mise au point de molécules actives, à un niveau symptomatique, sur des modèles cellulaires ou organiques, s'est donc renforcée d'une étape amont d'identification de cibles moléculaires précises. Ce sont ensuite sur ces cibles moléculaires précises que s'opérera la recherche et la mise au point de molécules actives.

Cette nouvelle étape amont d'identification des cibles fait l'objet de différentes approches. Certaines de ces approches

parmi les plus prometteuses sont caractérisées de « data driven » : il s'agit d'approches issues de **la confrontation des données cliniques et des phénotypes des patients avec des données de génotype, transcriptome et/ou protéome de ces mêmes patients (par exemple avec les profils d'expression génique fournis par les microarrays).**

L'objectif est de confronter et d'analyser ces grands volumes de données pour en extraire des hypothèses biologiques intéressantes concernant le mécanisme pathologique d'intérêt.

Un premier travail consiste à organiser ces données : il s'agit de mettre en place des solutions de data-management efficaces (tissuthèques, LIMS ...)

Un deuxième travail consiste à **traiter statistiquement** ces données biologiques avant de procéder à leur **annotation fonctionnelle** qui consiste à collecter toute l'information biologique les concernant.

Pour comprendre les **pathways**/mécanismes biologiques sous-jacents qui permettent d'interpréter les données analysées, nous nous appuyons ensuite sur LSGraph®, outil d'extraction d'informations à partir de la littérature. On obtient ainsi des cibles potentielles, sur lesquelles des études complémentaires peuvent être menées : analyses de séquences pour rattacher ces cibles d'intérêt à des familles plus larges ou pour déterminer leurs mécanismes de régulation, évaluation



J'attire votre attention sur le fait que les recherches en biologie et santé humaine mènent souvent à des plans expérimentaux compliqués où les modèles à appliquer ne sont pas forcément disponibles. La bioinformatique reste dans ce cas une question *d'hommes et d'expertises avant d'être une question d'outils disponibles.*

final sous une forme synthétique, comme si le biologiste n'avait en face de lui qu'une seule et même source d'information (voir figure 32).

Revenons sur l'étape « **Pathways** ». À partir des gènes d'intérêt identifiés, il est nécessaire de déterminer les mécanismes biologiques associés, qui expliquent les transitions de phénotype entre les différents échantillons du plan expérimental. L'objectif est d'établir une représentation de ces mécanismes, de sélectionner les points d'articulation et de découvrir de nouvelles cibles.

La solution retenue par IT.Omics est de s'appuyer sur des bases au contenu structuré contenant des données sur les mécanismes biologiques. Ces bases sont cependant très partielles. La principale source d'information reste la littérature scientifique dont l'inconvénient majeur

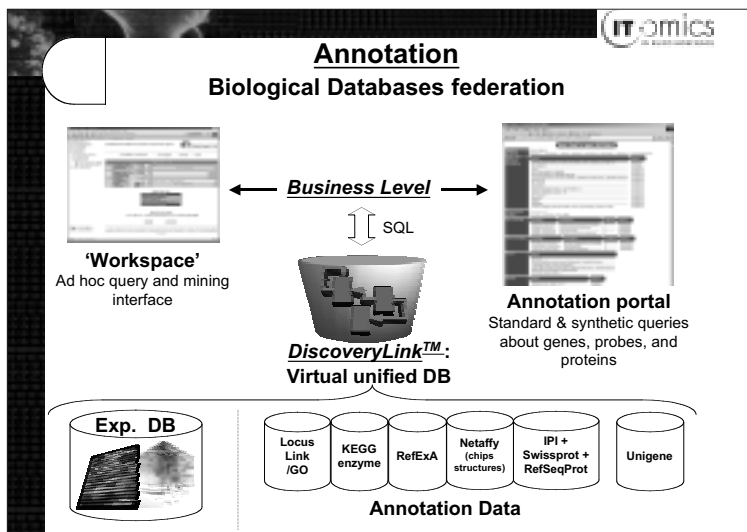


Figure 32

reste l'extraction et la structuration des données. La base de données Medline, par exemple, qui regroupe les abstracts scientifiques publiés dans les sciences de la vie, contient 13 millions d'entrées et augmente de 10 000 nouvelles entrées par semaine (voir figure 33).

LSGraph® extrait à partir de Medline l'ensemble des informations fonctionnelles décrites entre entités biologiques (gènes, protéines, etc.). Il représente ces interactions sous forme d'un graphe de connaissances et propose des outils permettant de

manipuler les données et d'en extraire de nouvelles connaissances intéressantes.

53

La première problématique dans cette stratégie est d'identifier les entités biologiques. Un gène ou une protéine ne sont pas des entités simples et sont souvent connus sous de nombreux synonymes. La seconde difficulté consiste à extraire de la littérature une information structurée, de façon automatique et la troisième enfin, consiste à représenter cette connaissance de manière utilisable par le biologiste (voir figure 34).

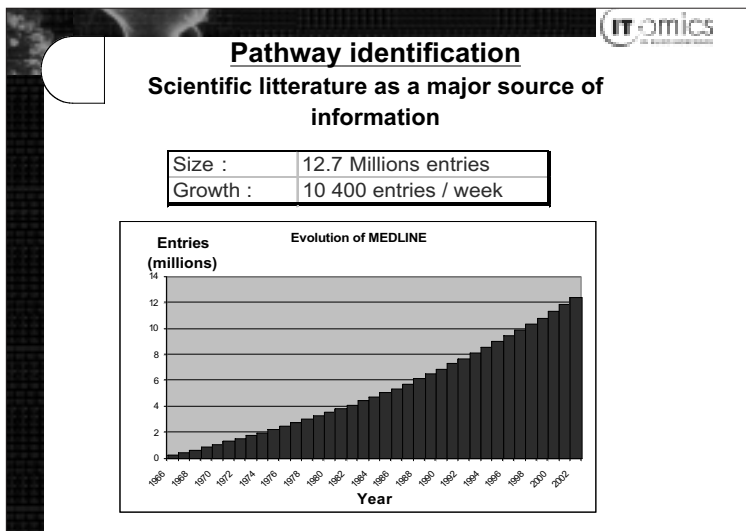


Figure 33

54

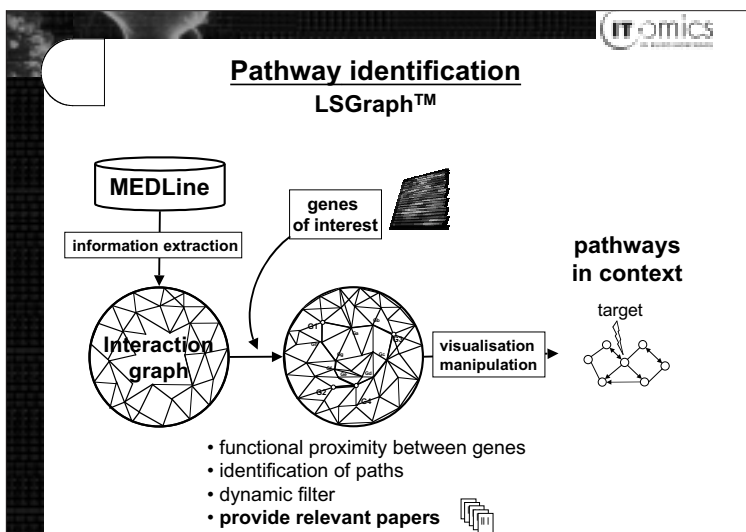


Figure 34

LSGraph® se présente sous la forme d'un graphe de connaissance métré qui décrit les relations entre les gènes au sein du « bibliome » (= la somme des articles scientifiques publiés à travers le monde). Il est possible avec LSGraph® de clusteriser ces gènes, de connaître leurs organisations et leurs proximités au sein de la littérature. On peut en extraire des sous-graphes ou revenir aux abstracts qui justifient les relations (voir figure 35).

L'un des grands problèmes actuels de la mise au point de médicaments est la compartimentation des différentes étapes du pipeline global que je viens de vous présenter. Ces étapes restent très cloisonnées entre elles.

L'un de nos objectifs aujourd'hui, au sein d'IT.Omics, est de faire en sorte que ce pipeline soit le moins séquentiel possible et que le biologiste, à tout instant, puisse simultanément adresser une question croisée à l'ensemble des sources d'information et des services disponibles. L'objectif d'IT.Omics est ainsi d'intégrer non seulement des données, à travers DiscoveryLink, mais aussi d'intégrer des informations plus larges, des données expérimentales, l'information de littérature, un certain nombre de services associés pour aider le biologiste dans l'identification de cibles prometteuses (voir figure 36).

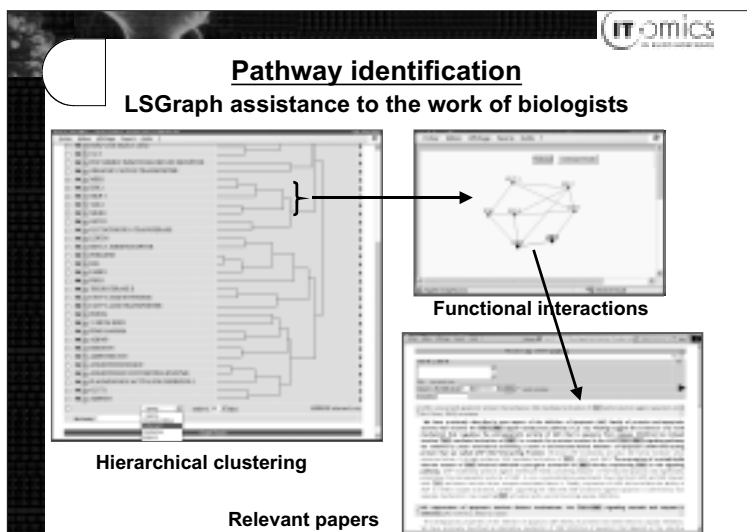


Figure 35

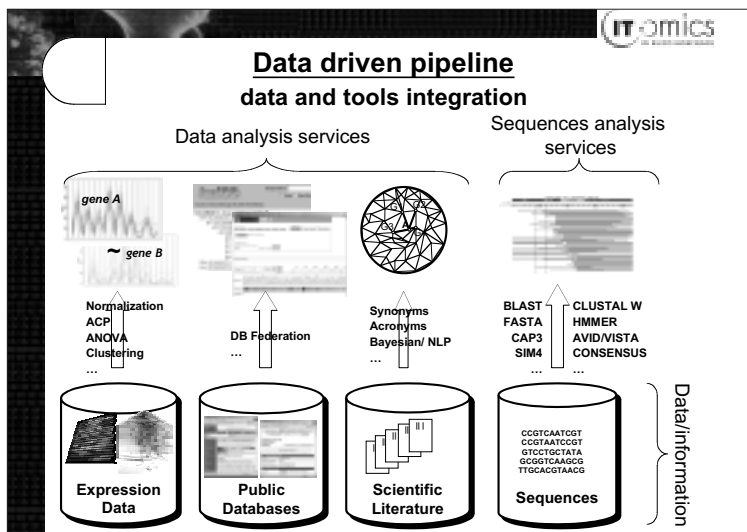


Figure 36



IPSOGEN
2 Boulevard Luce 13008 Marseille
Tél. +33 491 293 090
fert@ipsogen.com
www.ipsogen.com

Vincent Fert
CEO, Ipsogen

Oncogénomique, intégration des données cliniques et génomiques pour le traitement personnalisé des cancers.

Ipsogen est une jeune société de biotechnologie créée fin 1999, sur le concept de *la pharmacogénomique des cancers*. Nous travaillons sur *l'identification des marqueurs prédictifs de l'agressivité des tumeurs et de la réponse thérapeutique aux agents anti-cancer* à partir d'une plate-forme technologique de

profil d'expression de gènes à grande échelle. Nous sommes associés à des centres cliniques experts en oncologie.

Le traitement du cancer est aujourd'hui **séquentiel**. À une tumeur donnée correspondent un pronostic et un traitement standards. Lorsqu'un traitement échoue, il est remplacé par un traitement plus agressif et ainsi de suite, en sachant que les chances de survie et de réponse thérapeutique diminuent à chaque réorientation du traitement (voir figure 37).

57

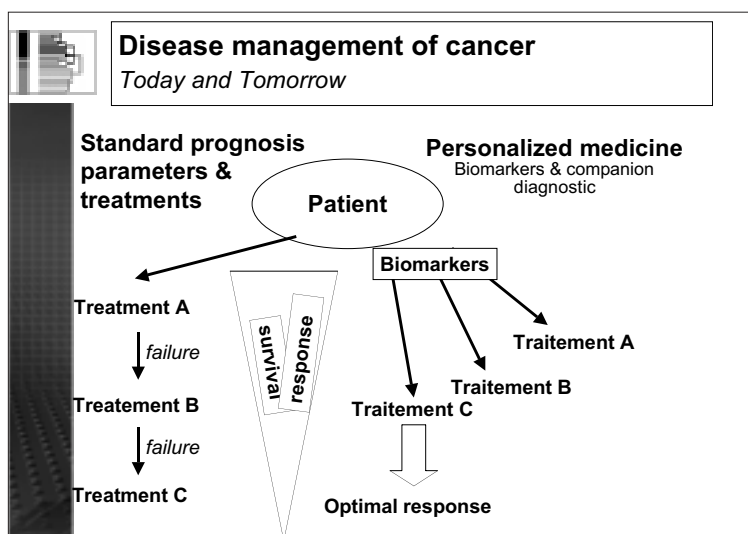


Figure 37

Grâce aux biotechnologies, nous progressons vers des traitements personnalisés : ***l'utilisation des biomarqueurs*** permet de sélectionner les patients afin d'optimiser la réponse thérapeutique. Certains médicaments en oncologie illustrent déjà l'efficacité de ce concept :

- Traitement à l'Herceptin développé par Genentech qui est un anticorps monoclonal administré aux patients exprimant le récepteur ERBB2, avec un test qui est utilisé dans les centres cliniques.
- la molécule Glivec (Novartis), qui s'adresse à une catégorie de patients atteints de leucémie et porteurs de la translocation BCR-Abl.

La plupart des sociétés impliquées dans le développement de molécules anti-cancer ciblées ***n'ont cependant pas ou peu intégré la prise en compte des biomarqueurs*** dans leur processus de recherche. Les résultats régulièrement négatifs des essais cliniques de ces traitements montrent d'ailleurs leur efficacité restreinte (par exemple, l'Avastin de Genentech) et les agences réglementaires prennent aujourd'hui position pour l'incorporation de données pharmacogénomiques dans le développement de médicaments.

La pharmacogénomique constitue un véritable ***défi scientifique*** et pose la question de l'existence, dans une tumeur primaire, de gènes prédictifs de ***l'agressivité de la maladie, de son évolution, et de la réponse à une thérapeutique donnée***. De telles signatures géniques existent dans les tumeurs primaires mais leur pertinence reste à démontrer sur de grandes séries de patients dans des essais multicentriques prospectifs.

Ce défi scientifique se double de plus d'un défi ***technologique*** puisqu'il s'agit de fédérer un ensemble de technologies et d'informations, pour réussir à transposer à l'oncologie ce concept de pharmacogénomique. Deux pré-requis sont nécessaires pour cela : d'une part, ***l'accès à des échantillons hautement qualifiés et aux informations cliniques qui leur correspondent***, et d'autre part, ***la maîtrise des technologies haut débit*** de mesure de profils d'expression.

Ipsogen a développé des partenariats étroits avec les centres anti-cancer afin de résoudre les problématiques liées au premier de ces deux pré-requis, et qui incluent :

- le contrôle de la qualité du stockage des échantillons qualifiés de tumeur pendant des périodes assez longues (essais prospectifs) et l'homogénéité des pratiques des chirurgiens et des pathologistes sur ces mêmes périodes.
- Concernant les annotations cliniques, l'adoption de critères d'inclusion stricts, garantissant notamment l'homogénéité des traitements appliqués. La connaissance de l'évolution de ceux-ci au cours de la maladie est également nécessaire pour pouvoir corrélérer la réponse thérapeutique et l'existence de marqueurs prédictifs.
- Des problèmes d'origine éthique : il s'agit de s'assurer que les données exploitées ont fait l'objet d'un consentement éclairé de la part des patients et respectent la loi Informatique et Libertés, dans le cas de leur incorporation au sein de bases de données.
- Des problèmes liés à la propriété intellectuelle et à son partage, dans le cadre de partenariats avec des instituts publics.

- Enfin, un des points clefs réside dans la capacité à fédérer des compétences hétérogènes autour d'un même projet : oncologues, chirurgiens, pathologistes, spécialistes en biologie moléculaire et biotechnologie.

Concernant *le deuxième pré-requis*, à savoir la maîtrise des technologies haut débit de mesure de profils d'expression, Ipsogen a développé une plate-forme, qui présente les caractéristiques suivantes :

- Analyses à haut débit : 120 échantillons par mois, en accord avec les bonnes pratiques de laboratoires (la validation est faite dans un contexte d'essai clinique et les données doivent pouvoir être incluses dans des dossiers d'enregistrement).

- Développement d'algorithmes spécifiques.
- Participation à la standardisation des structures d'échanges de ces données, au sein du consortium MGED (recommandations MIAME et MAGE-ML).

Nous avons développé une base de données (DiscoverySoftware) qui regroupe des mesures d'expression sur certaines tumeurs importantes, y compris des données d'immuno-histochimie à haut débit (voir figure 38).

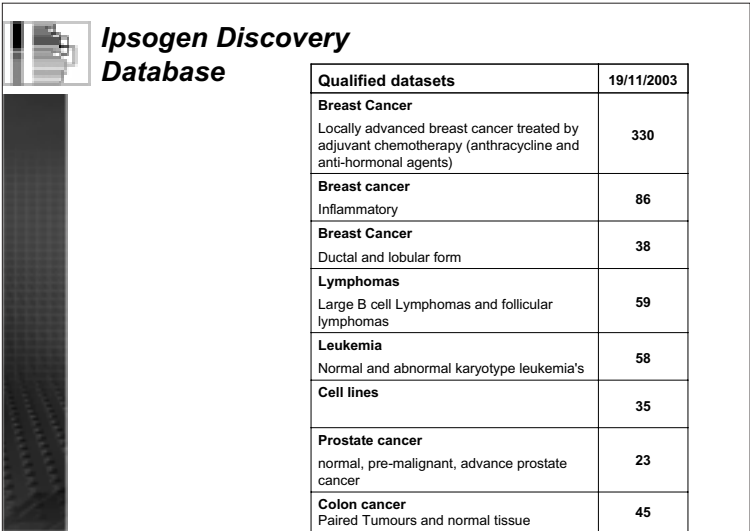


Figure 38

Cette plate-forme nous permet, à partir de puces développées en interne, d'effectuer du profiling à haut débit de l'expression de gènes sélectionnés pour leur intérêt en oncologie.

Nous avons également développé des outils pour **la gestion et l'analyse des données** :

Gestion des données (voir figure 39) :

- outils sous forme d'interfaces web pour assurer la traçabilité des expériences de

profils d'expression et permettre la validation des données, l'intégration des données cliniques.

- modules d'exploitation et d'extraction de bases de données génomiques existantes.

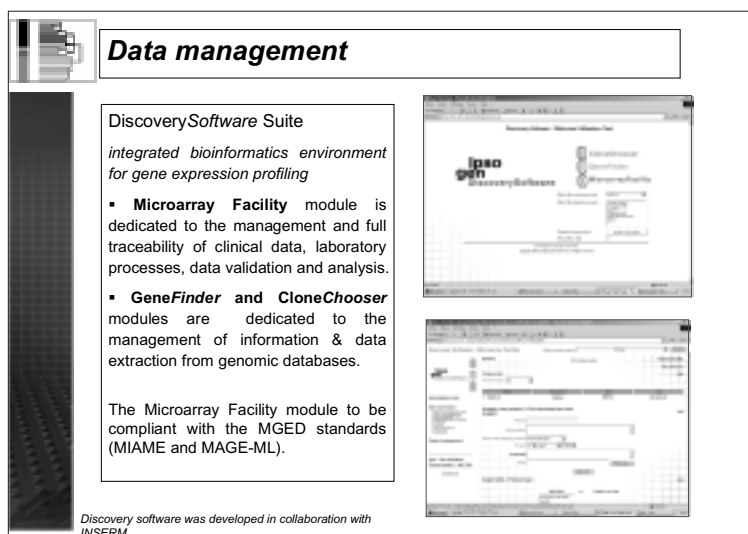


Figure 39

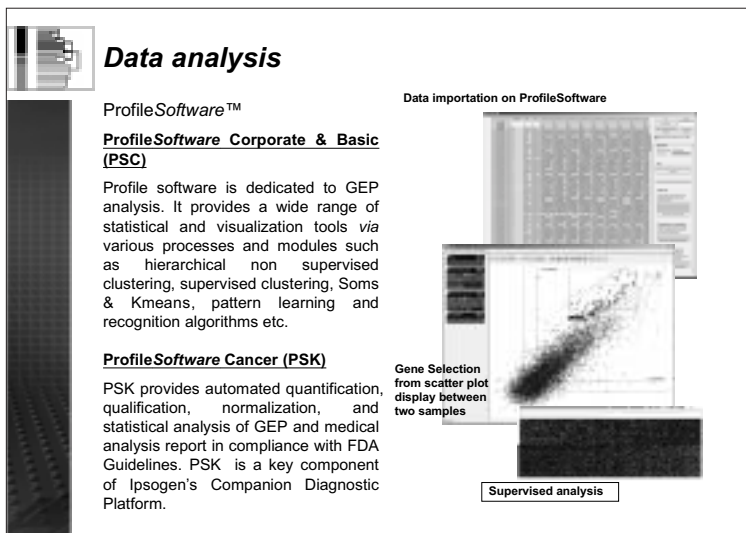


Figure 40

Analyse des données (voir figure 40) :

Nous avons développé un outil permettant la quantification, la qualification, la normalisation et l'analyse statistique des profils d'expression des gènes, outil fondé sur l'intégration des techniques de biostatistiques existantes.

L'ensemble de ces outils est disponible par le biais de prestations de services pour les industriels de la pharmacie ou des biotechnologies.

Ces bases de données et la propriété intellectuelle développée sur un ensemble des gènes prédictors de l'agressivité de certaines tumeurs constituent les principaux actifs de la société (voir figures 41 et 42).

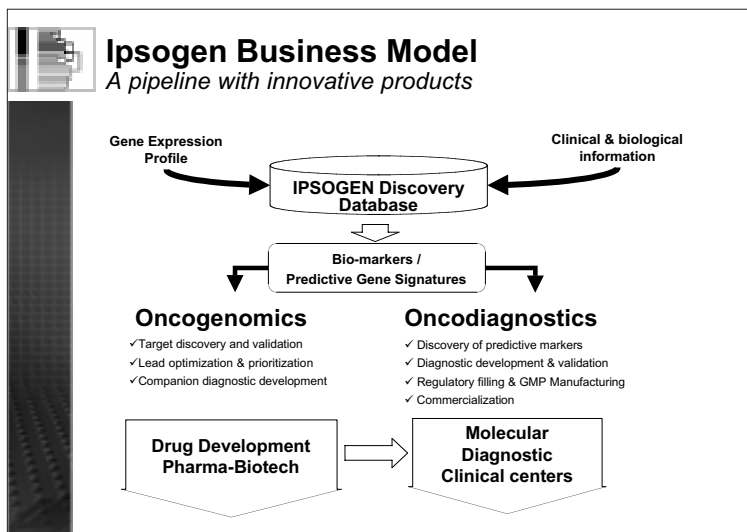


Figure 41

62



Figure 42

Clarisse Dhaenens
Laboratoire d'Informatique
Fondamentale de Lille,
Université de Lille I

ÉQUIPE OPTIMISATION PARALLÈLE ET COOPÉRATIVE
Bat M3 - Cité Scientifique
59655 Villeneuve d'Ascq cedex
Tél. 03 20 43 67 49 - Clarisse.Dhaenens@lilfl.fr
<http://www.lilfl.fr/~dhaenens>

Interprétation de données par méthode d'optimisation.

Notre équipe a été confrontée à la problématique du traitement des données génomiques dans le cadre de la Génopole de Lille consacrée essentiellement à l'étude de maladies multifactorielles telles que le diabète, l'obésité, etc. Plusieurs laboratoires se sont fédérés au sein de la Génopole pour étudier ces différentes maladies, regroupés en différents plateaux : séquençage et génotypage, protéomique, transcriptome et un plateau centré sur la bioinformatique : **le Centre Intégré de Bioinformatique**.

Ce plateau a deux missions :

- Mettre à disposition des biologistes et des chimistes un ensemble d'outils existants sur le marché (base de données d'expression issues de puces à ADN par exemple).
- Développer de nouveaux outils permettant des analyses plus poussées ou plus spécifiques en réponse à des problématiques particulières.

L'approche générale que nous utilisons consiste à **modéliser une problématique biologique**, pour en comprendre les différents termes, sous la forme d'une problématique **de fouille de données** qui recouvre différentes tâches :

- La sélection d'attributs,
- La segmentation (clustering),
- Les règles d'association,
- La classification.

Nous avons choisi d'assimiler cette problématique de fouille de données à une problématique **d'optimisation combinatoire** qui consiste à maximiser ou minimiser une fonction objectif dans un choix de solutions potentielles (voir figure 43).

Avant de donner un exemple d'étude, je vais détailler les **tâches de datamining** qui nous intéressent.

- La première tâche consiste à **réduire le nombre d'attributs**. Un attribut, par exemple un gène, caractérise une instance (un individu ou un couple d'individus que l'on compare). Les instances correspondent aux lignes de nos tables de données. L'objectif est de réduire le nombre d'attributs sélectionnés pour augmenter la précision de la prédiction et/ou réduire le temps de traitement des données.
- Le **clustering** (segmentation dans un cadre d'apprentissage non-supervisé) permet de regrouper des données similaires sans disposer forcément de notions de classes, c'est-à-dire sans savoir par rapport à quelles caractéristiques regrouper les données. Cette tâche implique une notion de distance ou de similarité.
- Les **règles d'association** constituent un modèle plus général dont l'objectif est de découvrir des motifs et corrélations, à partir d'ensembles d'items contenus dans des bases de données et de les exprimer de façon compréhensible.

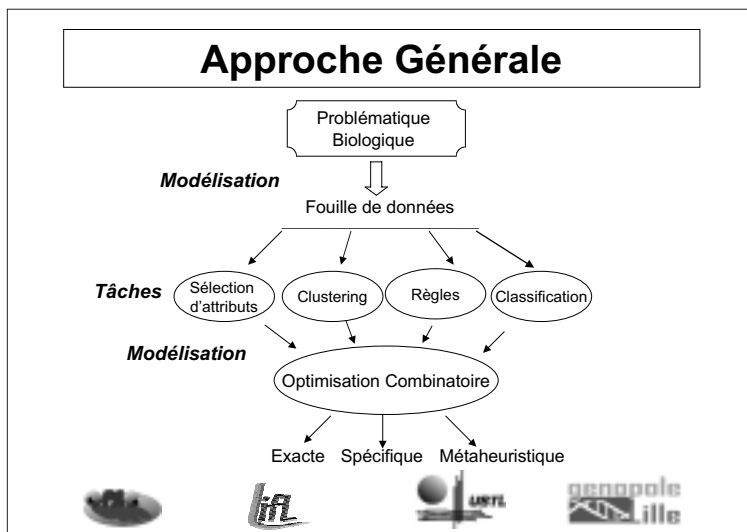


Figure 43

64

- La classification permet quant à elle de regrouper des données partageant une même caractéristique.

Je vais maintenant vous présenter deux études que nous avons menées : l'une concernant **la sélection d'attributs et le clustering dans le cadre d'une étude de paire de germains** et l'autre qui a eu recours à **une approche par règles d'association**. Notre angle d'attaque a été d'aborder ces problématiques grâce à une **approche par optimisation combinatoire**, qui est fonction du problème à résoudre (voir figure 44).

La première étude a été effectuée dans le cadre d'une collaboration avec le Laboratoire de Génétique des Maladies Multi-factorielles (LGMM) et l'Institut de Biologie de Lille (IBL), qui travaillent notamment sur le diabète et l'obésité. Les objectifs de cette étude étaient de repérer

des interactions et des cotransmissions de gènes pour **localiser un ou plusieurs gènes de prédisposition de ces maladies**.

Nous nous sommes intéressés aux **interactions génétiques et environnementales**. Pour une paire d'individus d'une même famille, nous avons comparé, par rapport au diabète, le partage de certains marqueurs génétiques et de facteurs environnementaux (âge, sexe, Body Mass Index, etc).

Nous étions confrontés aux problèmes suivants :

- Les hypothèses reposaient sur un nombre important d'attributs.
- Nous étions dans un contexte d'apprentissage non-supervisé, puisque la maladie peut ne pas être déclarée à l'instant où l'étude est effectuée. Le statut « non-malade » reste donc toujours incertain.

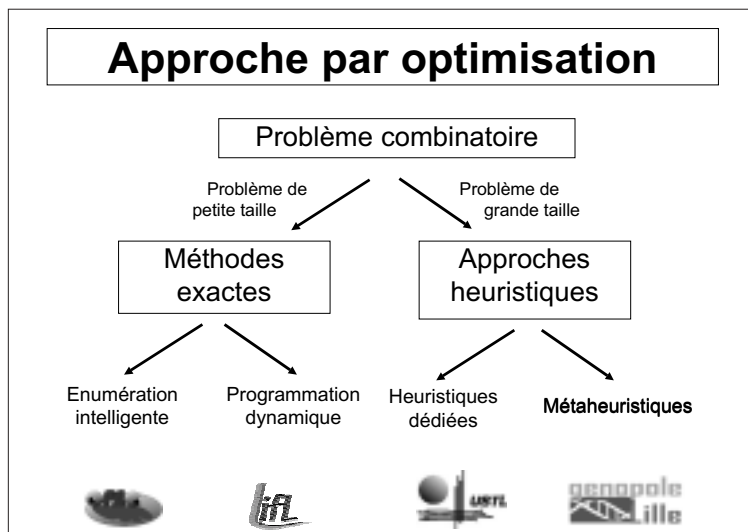


Figure 44

- Beaucoup de valeurs étaient manquantes, à cause des conditions d'expérimentation.

Notre objectif était de découvrir des groupes d'attributs responsables de la maladie.

Nous avons développé une approche en deux phases :

- Une sélection d'attributs qui permettait de fortement réduire le nombre de marqueurs considérés et de rendre possible une approche par clustering par une méta-heuristique (algorithme génétique).
- Une deuxième phase plus classique, le K-means, mais qui n'a pas donné de résultats satisfaisants. Nous avons en conséquence là aussi développé un algorithme génétique (voir figure 45).

La deuxième étude que je souhaitais vous présenter concerne **une approche par règles d'association**. L'intérêt d'utiliser

les règles d'association est lié au fait que nous avons à traiter des **données hétérogènes** (expérimentales, médicales, fonctionnelles) ; le nombre d'attributs s'avère très élevé (10 000 gènes en augmentation régulière) et nous recherchons des sous-groupes de gènes qui ont des comportements et des caractéristiques similaires. Ce problème général peut être facilement modélisé comme un problème de recherche de règles d'associations.

Une règle est une fonction qui a la forme « **si** (condition) **alors** (prédiction) » où la condition peut porter sur l'expression d'un gène (si tel gène est surexprimé et tel autre sous-exprimé, alors tel autre gène sera surexprimé). Certains critères existent pour mesurer la qualité des règles. Les deux principaux sont la **fréquence** (d'apparition de la règle dans la base de données) et la **confiance**, qui mesure la véracité de la règle (i.e. combien de fois se réalise la

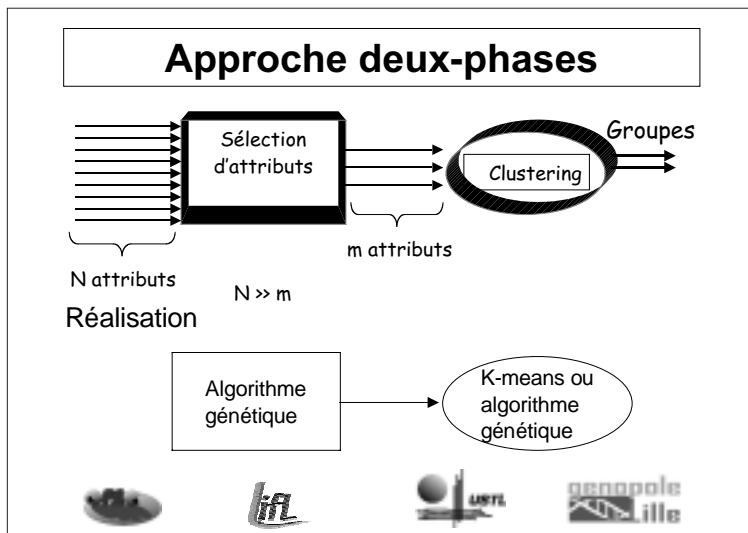


Figure 45

66 prédiction lorsque la condition est vérifiée) (voir figure 46).

Les règles d'association peuvent fournir différents résultats :

- Mettre en liaison des traitements utilisés sur des patients et la sur- ou sous-expression de certains gènes.
- Mettre en relation l'expression des gènes.
- Mettre en évidence des relations entre gènes dans un sous-ensemble de situations.
- Mettre en relation des comportements de gènes et d'autres données (caractéristiques fonctionnelles, structurales) (voir figures 47 et 48).

Présentation du modèle

Règle : R = IF C THEN P

Condition, partie gauche de la règle (C)

Conséquence, partie droite de la règle (P =Prédiction)

2 mesures principales de mesure de qualité

$$\text{Support}(R) = \frac{|C \& P|}{|BD|} \quad \text{\% de transactions qui contiennent simultanément C et P}$$

$$\text{Confiance}(R) = \frac{|C \& P|}{|C|} = \frac{\text{Support}(C \& P)}{\text{Support}(C)}$$



Figure 46

67

Intérêt de la modélisation

Règles d'association peuvent fournir différents résultats

✓ Traitement \Rightarrow Expression de certains gènes

Tr1 \Rightarrow {Gène A \uparrow , Gène B \downarrow } [Creighton - Hanash, 03]

Relations entre gènes (cas général)

{Gène A \uparrow , Gène B \downarrow , Gène C \uparrow } \Rightarrow Gène D \uparrow [Kotala et al, 01]

Relations entre gènes (pour certaines situations) [Becquet et al, 02]

{(Gène A \uparrow , Gène B \uparrow) dans situation y} \Rightarrow Gène D \uparrow dans situation y

Comportement de gènes \Rightarrow Caractéristiques fonctionnelles

\Rightarrow Caractéristiques structurales

{Gène \uparrow dans situation y} \Rightarrow Fonction X [OPAC, IT-Omics, 03]



Figure 47

Exemple concret

Action Bioingénierie – GenHomme Network (2001-2003)

“Data mining platform from heterogeneous data for cardiovascular diseases”,

IT-OMICS, GENFIT, LIFL

Recherche de marqueurs de prédisposition aux maladies cardiovasculaires

Expérimentations sur puces à ADN

- ✓ Puces Affymetrix
- ✓ 22 000 gènes étudiés
- ✓ patients à risque cardiovasculaire

Mise en correspondance de données hétérogènes

- ✓ Expérimentations sur puces à ADN
 - ✓ Données structurales (localisation cytogénétique, ...)
 - ✓ Données fonctionnelles (Groupe GO, Famille PFAM, ...)
- } BD publiques
GENBANK,



Figure 48

Discussion

Geneviève Martin

Quelle est la probabilité que vous tombiez sur une association, pur résultat de la statistique, qui soit fausse ?

Claire Dhaenens

Dans le cas des règles d'association, le critère est primordial : nous appliquons des méthodes qui vont optimiser ce critère. La fréquence d'une association peut par exemple être le seul fruit d'un calcul statistique. Notre approche est complétée par une approche multicritère. La collaboration avec les scientifiques est alors très importante pour le choix de ces critères.

Geneviève Martin

Existe-t-il des systèmes de classification internationaux communs pour uniformiser le nom des gènes ?

Guillaume Lièvre

Il existe un vrai problème autour de la notion de gène, qui évolue avec le temps. La nomenclature internationale Hugo permet d'attribuer à chaque gène suffisamment connu un nom unique, normalement utilisé par l'ensemble de la communauté scientifique. Ce n'est pas toujours le cas dans la pratique. Avant que le gène soit vraiment connu, plusieurs noms peuvent avoir été utilisés au niveau international. La gestion de l'historique de ces noms pose évidemment question. La réponse serait de construire au sein des organisations des référentiels solides. À un niveau plus global, il faut attendre que l'information tombe dans le domaine public, qu'un nom se dégage et que des outils informatiques puissent gérer l'historique, c'est à dire l'ensemble des docu-

ments où le gène a été décrit sous des noms divers et variés.



eurasanté

Entreprendre et réussir

EURASANTÉ

Parc Eurasanté

310 Avenue Eugène Avinée

59120 LOOS

FRANCE

Tél. (33) 3 28 55 90 60

Fax (33) 3 28 55 90 61

Site internet : <http://www.eurasante.com>

E-mail : contact@eurasante.com

Document réalisé grâce au concours de

