https://www.laurentbloch.net/BlogLB/38-L-Institut-Pasteur

L'Institut Pasteur

Les chats ne sont pas de gauche -

Date de mise en ligne : mardi 25 octobre 2022

Chapitre précédent[/Chapitre suivant/]

En 1991, alors que je suis au Cnam, une amie pasteurienne, Dominique Morello, me met en contact avec François Rougeon, Directeur de la recherche à l'Institut Pasteur, qui a conscience du retard que prend l'Institut dans le domaine de l'informatique scientifique, et nous commençons à réfléchir ensemble aux moyens de combler ce retard. À la fin de l'année je rends mon tablier au Cnam et rejoins l'Institut Pasteur pour y créer un Service d'informatique scientifique (SIS).

Le premier décembre 1991 j'arrive à l'Institut Pasteur pour y prendre mes nouvelles fonctions. Je connais déjà l'équipe en place parce que depuis plusieurs mois je travaille avec François Rougeon et je participe à des réunions pour déterminer un plan d'action informatique.

Tout est à faire. La situation est assez lamentable, avec les survivants d'une ancienne équipe d'informatique scientifique, que les personnels de statut chercheur, chef d'unité compris, ont quittée depuis deux ans. et au sein de laquelle le rang hiérarchique implicite est inversement proportionnel au contact réel avec l'informatique, puisque pour appartenir au clergé il faut pouvoir se dire, peu ou prou, « scientifique », c'est-à-dire biologiste [1]. Les informaticiens sont définitivement considérés comme du Tiers-État, ils essaient d'obtenir des cosignatures en mauvaise position pour les articles auxquels ils ont contribué et n'y parviennent que dans la mesure où cette contribution n'est pas de l'informatique.

L'infrastructure repose sur des ordinateurs Data General sous système AOS, ce qui n'a peut-être pas été un choix malheureux au début des années 1980, mais interdit l'accès aux logiciels les plus utiles et les plus désirés par les chercheurs.

L'outillage logiciel est essentiellement constitué du logiciel SASIP (Système d'Analyse de Séquences de l'Institut Pasteur), réalisé sur la base d'une adaptation au système AOS des programmes créés par Roger Staden (en 1982, en Fortran). Les banques d'acides nucléiques GenBank et EMBL occupent à l'époque chacune 150 mégaoctets (aujourd'hui plus de 16 pétaoctets), elles ne sont pas disponibles sur place, il faut les consulter par des interrogations à distance.

L'accès au réseau comporte un abonnement à Transpac et un raccordement à EARN/Bitnet par une ligne à 9600 bits/s. Il n'y a pas de réseau de campus mais une centaine de petits réseaux AppleTalk de laboratoire.

Bref, plus aucun de ces éléments n'est adapté aux perspectives de l'heure, et il n'est besoin de nulle hardiesse pour diagnostiquer que rien n'en est à garder.

De la communauté pasteurienne émane une demande diffuse de meilleurs moyens techniques et de plus d'assistance humaine pour « l'analyse de séquences ». Tout le monde se rend bien compte qu'il y a urgence, et même que l'on frise le ridicule si l'on se compare à l'état de l'art international. Il y a bien sûr des chercheurs parfaitement au fait de la réalité désignée par ces mots, mais pour d'autres c'est quelque-chose de mystérieux, rendu nécessaire par l'évolution des exigences des comités de lecture, mais qui ne fait pas vraiment partie de leur métier de biologistes. Il n'est pas inutile de rappeler l'état d'esprit de cette population mal informée.

Pour beaucoup, la vision de l'avenir meilleur attendu, où leurs séquences seraient analysées, ressemble à une salle de bureau de poste où derrière les guichets se tiendraient des ingénieurs compétents et disponibles. Les biologistes viendraient apporter leurs séquences sur des disquettes ou des bouts de papier, les donneraient aux ingénieurs qui

n'auraient plus qu'à « faire l'analyse ». D'ailleurs à la réflexion on pourrait imaginer de sonner les ingénieurs au téléphone, ce seraient eux qui se dérangeraient, ce serait mieux. Cette vision n'est pas une lubie : elle m'a souvent été dépeinte, et beaucoup de ceux à qui je l'ai proposée ironiquement m'ont fait part de leur enthousiasme et de leur désir de la voir réalisée.

Cette vision n'a jamais été réalisée et ne le sera jamais pour une foule de raisons parmi lesquelles deux ou trois que je vais exposer brièvement à l'intention du lecteur peu informé des réalités de la biologie moléculaire.

Le chercheur traditionnel en biologie moléculaire étudie, par exemple, un gène dans un organisme. Pour ce faire il effectue des manipulations longues et complexes (cela peut se compter en mois de travail) qui lui permettent d'obtenir des fragments d'ADN qu'il va séquencer. Tout ceci se passe à la paillasse (en « phase humide »).

Une fois la séquence obtenue, les logiciels de comparaison de séquences vont permettre de la comparer aux séquences des banques pour savoir si on ne vient pas de séquencer un fragment déjà connu du génome de l'organisme étudié, ou si ce fragment n'est pas similaire à un fragment déjà connu d'un autre organisme. D'autres programmes détectent les « régions codantes » et les gènes dans les séquences. Ce ne sont que des exemples, bien d'autres types d'analyses peuvent être entrepris.

Le problème est que ces analyses ne sont pas simples. Les séquences à comparer ne sont pas exactement égales, elles ont une certaine similarité qu'il s'agit d'apprécier au moyen de logiciels qui reposent sur des modèles probabilistes. Ceci se paramètre, et selon les valeurs des paramètres les résultats diffèrent. Bref, c'est un travail de recherche, et la recherche doit être faite par les chercheurs. L'idée de porter ses séquences à l'analyse comme ses chemises au pressing est anti-scientifique.

Frédérique Galisson, chercheuse biologiste dans l'unité de Génétique moléculaire des levure dirigée par Bernard Dujon avant de rejoindre le SIS, écrit dans le rapport d'évaluation de l'équipe : « Qu'il s'agisse de la prédiction de séquences codantes, de recherches d'éléments régulateurs dans le génome, de mesure de similitudes entre des séquences, de recherche de motifs particuliers etc., de nombreux programmes informatiques existent, qui produisent des solutions (d'exactitude et de validité variables) à des questions variées. Ces programmes mettent en œuvre des méthodes de calcul et reposent sur des modèles et hypothèses biologiques implicites lors de leur utilisation. Les méthodes de calcul peuvent être exactes ou heuristiques, et de nature purement algorithmique (reposant sur une formalisation précise du problème et de la procédure de calcul) ou probabiliste. Pour évaluer la signification d'un résultat, soupçonner l'existence de faux-positifs ou faux-négatifs, il est nécessaire de connaître les méthodes de calcul mises en œuvre par les programmes. La connaissance des modèles et hypothèses biologiques incorporés dans les programmes est fondamentale pour donner un sens biologique à l'interprétation des résultats. De façon à correctement choisir le programme approprié et les paramètres adéquats en réponse à un problème particulier dans un cadre scientifique précis, il importe donc de connaître parfaitement ces programmes. Il peut également être opportun de modifier ou d'adapter un programme en fonction de la problématique ou de l'organisme considéré. »

Une autre source de malentendu tient à l'estimation de la quantité de travail nécessaire à l'analyse classique d'une séquence : le biologiste qui n'a jamais fait lui-même une telle analyse croit volontiers que deux heures suffisent, là où souvent il faut trois semaines. Il est difficile de tomber d'accord à partir de telles divergences. Bien sûr une minorité de chercheurs déjà lancée dans l'analyse de séquences avait d'autres exigences : disposer des logiciels classiques (à l'époque surtout Staden et GCG), avoir des mises à jour plus fréquentes des banques. Mais l'ampleur des malentendus relatifs à la nature du travail d'analyse de séquences et à la nature de l'informatique montrent qu'aux missions évoquées plus haut s'en ajoutera une autre, préalable : un travail d'éducation culturelle pour convaincre les Pasteuriens de l'existence de l'informatique et leur procurer quelques lumières sur sa nature et ses pouvoirs.

Mais aussi, si l'on veut que les biologistes fassent eux-mêmes leurs analyses de séquences, encore faut-il leur

donner les moyens de le faire, notamment des logiciels adaptés à l'usage d'un être humain normal, ce qui est loin d'être le cas général pour les programmes issus de la recherche. D'où l'importance du travail signalé déjà de spécification et de développement d'interfaces.

Mais, avant même cela, bien peu imaginaient l'ampleur et la complexité du travail à accomplir pour créer la condition préalable à toute installation et utilisation de banques de séquences et de logiciels d'analyse et à toute éducation informatique : doter l'Institut Pasteur d'un système informatique, d'un réseau de campus et d'un accès à l'Internet dignes de ce nom.

Dès les premières études de 1991 il était clair que donner à tous les chercheurs l'accès aux banques de séquences et aux logiciels d'analyse ne pouvait se faire sans la constitution d'un réseau de campus qui les relierait aux gros serveurs nécessaires aux données et aux traitements. Les ordinateurs personnels de l'époque n'étaient pas assez puissants pour effectuer ces calculs, il n'est pas certain qu'ils le soient aujourd'hui, et de toute façon les banques de séquences sont trop grosses, leur mise à jour est un processus trop complexe pour que chaque chercheur ait à s'en débrouiller.

Le choix s'était vite imposé d'ordinateurs dotés du système d'exploitation Unix. Beaucoup de Pasteuriens ont souvent au cours des années suivantes eu l'occasion de se plaindre de l'inconfort de ce système pour les utilisateurs, et non sans quelques raisons : Unix a été conçu et réalisé pour des chercheurs en informatique et des professionnels qui s'en servent en permanence et dont c'est le métier, pas pour des biologistes. D'un autre côté, le marché des systèmes Unix offre les machines les plus puissantes aux meilleurs prix et la plupart des logiciels scientifiques sont développés sous Unix, ce qui fait deux arguments non négligeables. Et aussi, tous les Pasteuriens que j'ai rencontrés à l'époque voulaient Unix à l'exclusion de toute autre chose, sans trop savoir pourquoi, mais c'était dans l'air.

Au printemps 1992 sont livrés un serveur Sun et un calculateur vectoriel Convex, tous les deux sous Unix. C'était une augmentation de deux ordres de grandeur de la puissance de calcul disponible pour la recherche à l'Institut Pasteur.

Une fois ces machines livrées, il a fallu y installer les premiers logiciels, les raccorder à l'Internet, créer les comptes des utilisateurs. Seuls ceux qui ne l'ont jamais fait croient que c'est facile et rapide. L'équipe en place était dépourvue d'expérience Unix et Réseaux. Et puis, faire marcher un système utilisé par des centaines de personnes n'a rien à voir avec la gestion d'un Macintosh personnel (ceci dit, si l'on comptabilisait les heures passées par chacun à installer et configurer des choses sur son Macintosh on serait en présence d'un important gisement de productivité inexploité).

Heureusement le SIS a pu renforcer ses rangs de personnes déjà formées : Frédéric Chauveau en avril 1992, Louis Jones en octobre 1992, Catherine Letondal en juin 1993 et Christophe Wolfhugel en juin 1994 constituent la première vague d'arrivées consécutives à la réorganisation. Il est à noter que tous sont des informaticiens professionnels confirmés, ce qui semble avoir surpris certains Pasteuriens. Pour appliquer l'informatique à la biologie il faut *faire de l'informatique*, et les gens qui savent faire ça sont les informaticiens.

Pour la construction du réseau la date était favorable : il était déjà clair que les technologies à base de câble coaxial étaient en déclin, et le couple fibre optique - paire torsadée était suffisamment mûr pour être stable et bon marché tout en étant suffisamment jeune pour avoir de l'avenir.

Ce changement de technologie allait de pair avec un changement de topologie : aux structures en bus (toutes les machines connectées partagent un même support de communication) succédaient les structures en étoile ou en

arbre. En apparence ces nouvelles structures sont moins intéressantes parce que moins denses. En fait elles sont (sur un réseau local où la qualité de transmission est uniforme et bonne) les plus sûres parce qu'elles évitent le risque de formation de cycles, que le dysfonctionnement d'un sommet ne perturbe que lui-même et ses éventuels descendants, que la localisation des perturbations est (relativement) simple et que la portée des diffusions est facile à limiter.

Pour la conception de ce réseau nous nous sommes inspirés de la réalisation, à l'époque récente, de l'Université Carnegie-Mellon de Pittsburgh. Une boucle en fibre optique parcourt le campus en profitant des souterrains omniprésents et dessert le pied de chacun des dix-huit bâtiments principaux par un premier niveau de répartiteurs. De chacun de ces répartiteurs part une colonne en fibre optique qui dessert les étages du bâtiment et éventuellement les plus petits bâtiments, en cascade. À différents étages dans les bâtiments se trouvent les répartiteurs de second niveau qui abritent des routeurs associés à des répéteurs (hubs, souvent remplacés maintenant par des commutateurs) d'où part le câblage capillaire en paire torsadée. Au total 4 km de fibre optique et 140 km de paire torsadée pour desservir 2200 prises dédoublables. Les travaux ont duré 9 mois et il y a eu jusqu'à 50 techniciens et monteurs sur le chantier. Les Pasteuriens ont supporté avec une patience remarquable la perturbation de leur travail et l'occupation par des routeurs et des platines de jarretièrage d'emplacements auparavant dévolus aux centrifugeuses et aux congélateurs.

Un réseau ce n'est pas seulement du câble, c'est aussi des équipements actifs et du logiciel, et pas des plus simples. Côté équipements actifs, l'épine dorsale était exploitée en Ethernet. Sur le câblage capillaire il fallait construire deux types de réseaux : Ethernet et, pour interconnecter les quelques cent-vingt réseaux de Macintosh présents à l'époque sur le campus, LocalTalk ou sa variante Phonenet. Pour ce qui est du logiciel, plus précisément des protocoles, on avait TCP/IP et Appletalk. Le réseau est divisé en sous-réseaux routés statiquement, ce qui permet d'éviter la propagation des incidents de fonctionnement.

Le partitionnement en sous-réseaux routés a un inconvénient : sa mise en œuvre et son administration nécessitent des compétences pour établir le plan d'adressage et le paramétrage des équipements actifs. En contrepartie, les avantages sont une sécurité accrue tant face aux pannes que face aux fausses manœuvres et aux tentatives d'intrusion, et un trafic réduit sur la plupart des segments. Typiquement, il y a du trafic vers les extrémités, entre les machines d'un même laboratoire, il y en a à la racine, là où se concentrent les accès aux serveurs et au point de passage vers l'Internet, et il y en a peu sur les parties intermédiaires du réseau.

En chiffres ronds, l'équipement initial du SIS en serveurs, stations de travail et logiciels a coûté quatre millions de francs de 1992. La construction du réseau a coûté douze millions de francs, qui se répartissent en quatre millions pour la pose de câbles et huit millions pour les équipements actifs, logiciels et systèmes de contrôle. De cette dernière somme de huit millions, quatre ont été couverts par une donation de *Digital Equipment*.

Une chose que nous ne voulions pas faire, et cela nous a été reproché mais tant pis : nous ne sommes pas un service de « micro-informatique » et n'avons pas vocation à assister les chercheurs dans l'usage de leur Macintosh ou de leur PC. Cette tâche, si tant est que c'en soit une, absorberait toute notre énergie, et elle est incompatible avec le développement d'une informatique de pointe pour la recherche. On ne peut pas demander aux mêmes gens de développer du logiciel scientifique, activité qui demande de la concentration et du temps ininterrompu, et de répondre au téléphone à toutes les personnes qui ne savent pas lire le mode d'emploi du traitement de texte qu'elles viennent d'acheter.

La demande d'assistance à l'utilisation de micro-ordinateurs est potentiellement infinie, et le travail pour la satisfaire est totalement dépourvu d'intérêt. Une personne qui se consacrerait à ce travail à plein temps serait, au bout de quelques années, techniquement déqualifiée. Mais c'est une demande récurrente de chercheurs haut gradés et vieillissants, qui finiront au bout de dix ans par obtenir mon départ.

Une fois cette infrastructure mise en place, l'équipe du SIS a pu s'attaquer à l'installation des banques de données biologiques et des logiciels destinés à leur exploitation. En effet, pour comparer une séquence biologique à une banque, il est possible de l'envoyer par courrier électronique au *National Center for Biotechnology Information* (NCBI, à Bethesda dans la banlieue de Washington), qui répond très rapidement. Cette façon de procéder présente des inconvénients : d'abord cela ne convient pas si on veut analyser de grands volumes de données. Ensuite, les experts du NCBI, qui sont les meilleurs au monde, peuvent suivre pas à pas la démarche de la recherche, ce qui n'est pas forcément une bonne idée dans l'univers compétitif de la recherche en biologie moléculaire, d'autant plus que ses applications industrielles peuvent constituer des enjeux économiques considérables. Bref, pour les données et les logiciels aussi, nous aurons nos propres infrastructures.

Ce n'est pas tout d'avoir des ordinateurs, des banques de données et des logiciels, il faut aussi apprendre aux gens à s'en servir. Très vite le SIS aura une activité de formation importante, animée principalement par Frédérique Galisson. Assez vite il apparaît qu'il ne suffit pas d'apprendre aux chercheurs à utiliser les banques de données et les logiciels existants : cette compétence peut suffire à des techniciens qui effectuent un travail répétitif, mais un chercheur doit comprendre en profondeur le fonctionnement des logiciels qu'il utilise, ne serait-ce que pour en régler les paramètres conformément au travail entrepris. Et pour cela il n'y a pas plusieurs moyens : il faut apprendre l'informatique.

Lors de mon arrivée à l'Institut Pasteur j'ai été reçu par son Directeur général, Maxime Schwartz, qui m'a conseillé de prendre contact avec un jeune chercheur de l'Unité de programmation moléculaire et de toxicologie génétique dirigée par Maurice Hofnung, William Saurin. William, très tôt, a compris le rôle que l'informatique allait jouer dans la recherche en biologie, il a eu des échanges intenses avec Pierre Sonigo, un des auteurs de la publication initiale de la séquence nucléotidique du virus du Sida (VIH), il s'est mis à la programmation et à l'informatique en général. Nous avons de nombreuses conversations, il m'apprend beaucoup de choses sur la biologie et ses rapports avec l'informatique, ainsi que sur l'organisation de la recherche, à Pasteur et dans le monde. C'est lui qui débarquera un jour dans mon bureau pour me proposer la création d'un cours d'informatique pour les Pasteuriens. Ce sera l'objet d'un prochain chapitre.

[/Chapitre suivant/]

---

[1] Il m'aura fallu six ans et un voyage à l'Institut Pasteur de Madagascar pour comprendre que dans le langage pasteurien traditionnel « scientifique » s'oppose en fait à « médecin ». Dans le cas qui nous occupe il s'agissait plutôt de l'opposition « noble » - « commun ».