Je vous parle d'un temps ..

Alain G.

guenoche.alain@free.fr

Pour comprendre la genèse du groupe de travail *Alignement et Phylogénie*, il faut reprendre l'histoire un peu avant sa création. Il faut partir de la première réunion *Génome et Informatique*, organisée par Alain Hainaut à Massy-Palaiseau, du 8 au 12 Avril 1991. J'ai eu la chance d'y participer et cela a complètement réorienté mes recherches.

Dans un hôtel de lointaine banlieue parisienne, afin que les participants restent sur place et dînent ensemble, il avait réuni de grands noms de la Biologie Moléculaire et des responsables d'équipes de Mathématique ou d'Informatique (ce que je n'étais pas vraiment). Le but était d'inciter les seconds à former des étudiants et à faire passer des thèses sur les sujets proposés par les premiers, car ils avaient bien compris que leur discipline évoluant rapidement ils avaient et auraient besoin de développements méthodologiques qu'ils étaient incapables de fournir par eux-mêmes.

Naïf parmi les naïfs, mais pas tout à fait car j'avais déjà collaboré avec une équipe marseillaise impliquée dans le séquençage de *Basilus subtilis* (Jacques Haïech, François Denisot), j'ai été initié aux grands problèmes de la Biologie Moléculaire qui nécessitaient de nouvelles méthodes informatiques et donc des études mathématiques. Je me souviens de remarquables leçons sur les difficultés du séquençage, de la comparaison des nouvelles séquences avec ce qui commençait à être accessible dans des bases de données, l'organisation de ces bases, de la recherche de gènes (parties codantes), des parties communes ou voisines avec d'autres protéines, de leur structures binaires des ARN et des repliements tertiaires prédictibles sans passer par la cristallographie, des interactions potentielles entre ces protéines et bien sûr de leur fonction au sein de la cellule. J'allais oublier l'évolution et la fameuse reconstruction phylogénétique qui nous intéresse toujours ici. C'est ce dont je me souviens des problèmes de l'époque. J'aimerais d'ailleurs retrouver le programme de cette réunion qui a relancé, en France, le développement de l'*Info-Bio-Math* comme j'ai toujours souhaité qu'on l'appelle.

Il en a découlé une série de réunions *Génome et Informatique* qui regroupaient toutes les équipes françaises travaillant sur ce thème, quelque soit le sujet. S'y retrouvaient donc des chercheurs venant de Paris (Pasteur, Orsay, Jussieu), Lyon, Grenoble, Marseille, Montpellier, Toulouse, Bordeaux, Lille, un vrai tour de France qui, comme toujours, finissait à Paris.

C'est à la seconde, les 4 et 5 Février 1992, que nous prîmes la décision de former des groupes de travail sur des sujets spécifiques. A la sortie, dans les couloirs du LAFORIA à Jussieu, Olivier Gascuel, Manolo Gouy et moi même décidâmes de lancer le groupe Alphy. Pourquoi nous ? Manolo G., parce qu'il dirigeait à Lyon, dans le laboratoire LBBE de Christian Gautier,

une équipe très productive sur le sujet ; c'était le seul à s'y connaître vraiment. Olivier G. parce qu'il avait déjà travaillé en bio-info (avec Antoine Danchin) et qu'il était vivement intéressé par le sujet, comme on a pu le vérifier par la suite. Et moi parce que j'avais co-écrit un livre *Les arbres et les représentations des proximités*, Masson 1988, qui portait sur la reconstruction d'arbres et leur tracé automatique. En fait il n'y a pas un mot sur l'évolution dans ce livre, mais il y a beaucoup de méthodes informatiques, en particulier pour la reconstruction d'arbres à partir d'une mesure de distance. Et Olivier et moi collaborions depuis plusieurs années sur des questions d'Apprentissage (méthodes symboliques-numériques). Peu après (Avril 1992), Antoine Danchin et François Rechenmann ont obtenu la création d'un Groupe de Recherche (GDR-CNRS) et Alphy y était référencé.

Le nom d'Alphy nous est venu spontanément, tant les deux problèmes, d'alignement et de phylogénie étaient indissociables pour identifier les mutations. Il faut rappeler qu'à cette époque, déterminer une séquence était une tâche ardue : il fallait la découper en courts fragments (d'au plus 400 bases) puis les cloner, décrypter des gels d'électrophorèse, lire les séquences (à la main) et les assembler dans des contigs par recherche de chevauchements. La séquence progressait à raison de 4 kb par jour. Le premier génome animal complet, celui du nématode C. elegans (97 Mb), commencé en 1989, à finalement été publié fin 1998. Donc les chercheurs les plus actifs travaillaient sur des fragments, des gènes ou de courtes séquences d'ARN ribosomiques, le plus souvent alignées à la main (Richard Christen avait ainsi réalisé une base de plus d'un millier de 16S de bactéries).

Les premières réunions

Faut s'y mettre : la première réunion eut lieu à Marseille les 14 et 15 Septembre 1993, c'est pourquoi on peut parler des 30 ans d'Alphy. Je n'ai pratiquement aucun souvenir du nombre de participants, des exposés, en dehors du mien sur l'ajustement des longueurs d'arêtes d'un arbre hiérarchique. J'ai noté qu'il y eut un diner collectif et que nous avons fini en développant un programme de travail.

A l'époque, pour l'alignement, le programme Clustal (Higgins & Sharp, 1988) pour deux ou de multiples séquences et, pour la reconstruction, la méthode de Neighboor Joining (NJ, Saitou, Nei, 1987) faisaient largement autorité et l'une des questions primordiales était de faire mieux. Mais comment définir une distance évolutive ? Largement utilisées aussi, les méthodes de parcimonie, mais qui nécessitaient de distinguer les *vrais* événements évolutifs, simples mutations de bases ou de suites de bases, impliquant ou non des changements

de codons, ou des sauts plus ou moins éloignés, etc..

La seconde réunion eut lieu à Montpellier les 19-20 Mai 1994 pour laquelle je n'ai pas plus d'information, sauf qu'elle se déroulait certainement au LIRMM, où vous avez peut être plus d'information. Elle venait une semaine après une réunion du GDR à l'Institut Pasteur à Paris qui avait invité Samuel Karlin, célèbre statisticien impliqué dans le fameux programme Blast de recherche de séquences voisines, pour un exposé : *New methods in molecular evolutionary comparisions*.

La troisième réunion en 1995 eut lieu à Orsay, organisée par Hervé Le Guyader qui avait participé à la toute récente rénovation de la galerie de l'évolution du Muséum d'Histoire Naturelle au jardin des plantes. Il nous y a emmené le 11 Mai au soir, alors que la galerie nous y était réservée et, tout en nous commentant la visite, il nous a raconté de piquantes anecdotes sur l'inauguration en compagnie de François Mitterrand qui, ne pouvant profiter du buffet, éprouvait un malin plaisir à faire durer la cérémonie !

A partir de cette date, nous avons tourné régulièrement, à Montpellier en 1996 (13-14 Juin), Lyon en 1997 (22-23 Mai), avec un exposé (marginal) de la controversée sociobiologie selon Edward Osborn Wilson (qui voulait justifier les pires comportements humains comme des héritages évolutifs), à l'Institut Pasteur en 1998 (8-9 Avril) sous l'impulsion de Marie-France Sagot qui venait d'y être nommée. Retour à Lyon en 1999 (11-12 Février), le lendemain du jour où nous avons commencé et achevé, avec Laurent Duret, un programme de reconstruction à partir d'une distance partielle (avec des valeurs manquantes) et qu'il avait activé le jour même ajoutant des centaines d'arbres à la base du LBBE. Un choc pour moi qui ne les calculait et traçait qu'un à un.

En 2000, nous sommes revenus à Montpellier (3-5 Mai) où Mike Steel était invité, bientôt auréolé de son impeccable ouvrage (*Phylogenetics*, 2003). En 2001, je revenais de mon long périple en bateau et Richard Christen a pris les choses en main. Il nous a réunis à Avignon les 27 et 28 Mars, puis ce fut à Lyon les 29-30 Avril 2002, et à l'Institut Poincaré les 16-19 Juin 2003, où nous avions invité Joseph Felsenstein, auteur du fameux package PHYLIP. Ce fut aussi l'année où Jobim fut confondu avec L'European Congres (ECCP) à la cité des Sciences à La Villette.

Je dois aussi mentionner les nombreuses réunions ou colloques qui ont eu pour thème la phylogénie. Dans tous les Jobim, il y avait une session dédiée, mais dès 1994, je peux citer, à Lyon les journées Jacques Cartier, une école CNRS sur le Traitement de l'information génétique à Asnelles/mer (qui nous a permis d'admirer la tapisserie de Bayeux), des groupes Evolution Biologique à Marseille plusieurs années durant, des journées Génomique et optimisation à Nice, une réunion "Phylogénie et évolution correllées" à l'Institut Poincaré

en 2000 ; j'en passe beaucoup. Tout cela pour souligner que nous sommes en moins de 10 ans devenus une communauté très active et très forte, bien plus que celles engendrées par les autres groupes de travail du GDR. Ma dernière liste de diffusion de ces années là contenait 81 adresses mail, ce qui constituait une part importante de la communauté Bioinfo.

Enfin vint 2004 et la réunion de Lyon les 15-16 Janvier. Lors du diner collectif à la Brasserie Georges, j'ai facilement convaincu Laurent Duret et Nicolas Galtier de reprendre les rennes et d'assurer la continuité d'Alphy, ce qu'ils ont très gentiment accepté. La charge n'était pas très lourde, mais il faut quand même s'occuper du budget (nous sommes partis de rien, mais avons fini par défrayer les doctorants), donc des subventions et bien sûr des rapports d'activités pour les justifier. J'ai perdu beaucoup de ces rapports, faute de changements d'ordinateurs et de logiciels, mais ceux que j'ai retrouvés (année 2001-2004) mentionnent la liste des exposés et un résumé quand il m'a été transmis. A partir de 2004, je n'ai plus rien excepté les dates et les lieux.

La SFBI

Comme excuse à mon détachement, je peux rappeler que j'ai créé, le 1 Juin 2005, la Société Française de Bioinformatique, suite à des échanges épistolaires avec quelques collègues qui ont bien voulu me suivre dans cette aventure et qui ont constitué le premier Conseil de la SFBI composé de : Jean Lobry et Claude Thermes (vice présidents), Joël Pothier (trésorier), Laurent Mouchard (webmaître), Nicolas Galtier et Jacques Nicolas (Secrétaires), Hidde de Jong et Yves Quentin (rédacteurs). Et je me suis résigné à être le président, pour la seule première année, car au renouvellement de 2006, j'ai cédé bien volontiers ma place à Guy Perrière. Il a été suivi de très actives présidentes, à commencer par Sophie Schbath. Au départ, le rôle le plus important était celui de webmaître et heureusement que Laurent Mouchard, qui s'occupait déjà de la liste bioinfo, a bien voulu réaliser un premier site, car j'étais incapable de le faire (et je le suis toujours).

La raison d'être de cette société savante, comme association "Loi de 1901", était la désignation du groupe qui prend en charge l'organisation de la conférence nationale, JoBim qui avait été instaurée par Olivier Gascuel et Marie-France Sagot en 2000. Mais pas seulement, car il fallait

- effectuer le recensement des équipes constituées,
- faire connaitre les enseignements universitaires et les écoles d'été,
- diffuser les postes mis aux concours CNRS, INRA, INRIA, INSERM, IRD, Universités, postdoc, etc..
- faire connaitre les appels d'offre des contrats de recherche,

— maintenir le répertoire des thèses et habilitations.

Toutes ces rubriques devaient être accessibles sur le site, ce qui fût le cas pendant plusieurs années.

Je me suis personnellement occupé de ce répertoire et je demandais aux nouveaux docteurs de m'envoyer un titre, un résumé et la composition du jury (avec mention des directeurs, rapporteurs et examinateurs). Je les ai retrouvé pour les années 2005 à 2010 et, en les feuilletant, j'ai dénombré respectivement 13, 37, 27, 54, 29 et 40 notices. Elles ne sont pas toutes en phylogénie, mais si quelqu'un veut s'y plonger, je peux les lui transmettre. Malheureusement, je n'ai pas de liste de thèses en phylogénie avant ces dates. Il y en eu bien sûr, ne serait-ce que la vingtaine dont j'ai été rapporteur (jusqu'en 2012).

Si j'ai voulu évoquer la génèse de la SFBI dans cette note historique sur Alphy, c'est pour souligner la forte influence des membres du groupe de travail sur la création et le fonctionnement de l'association.

Evolution thématique

En 1993, Alain Hainaut s'était taillé un beau succès en montrant la position des codons stop sur un chromosome de la levure. Et donc entre ces codons se trouvaient les gènes auxquels il suffisait d'appliquer le code génétique pour caractériser les protéines. Déduction sans doute bien naïve aujourd'hui. Mais c'est bien ça le sujet intéressant : comment la discipline a évolué. Les thèmes abordés en 1992 ont suivi leurs voies et je serais bien en peine de les retracer tous. Mais essayons pour les questions relatives à l'évolution et j'espère que parmi vous me viendront quelques détails supplémentaires et aussi quelques corrections.

Je ne parlerai pas, faute de compétence, de l'accroissement phénoménal du nombre de séquences dû à l'acquisition relativement facile de robots séquenceurs. Fini les découpes aléatoires, les gels d'électrophorèse et les lectures manuelles et multiples pour éviter les erreurs. Puis rapidement on a eu accès aux génomes complets. D'abord des centaines de bactéries qui ont permis d'étudier l'origine du vivant, jusqu'à plusieurs spécimens de la même espèce pour quantifier la diversité génétique. Mais revoyons quelques méthodes

Les méthodes de distance produisent des arbres non enracinés et l'on essayait de placer une racine à l'aide d'un "outgroup", une espèce dont on était sûr de l'antériorité. Mais suivant l'espèce choisie, l'arbre n'était pas toujours le même ! D'ailleurs, suivant l'espèce sélectionnée pour représenter un clade, ils ne s'organisaient plus de la même façon !

Puis est apparue la subtile distinction (pour les informaticiens) entre les gènes paralogues et orthologues, ceux qui dérivaient bien du même gène ancestral et non pas d'une duplication. D'où la recherche des gènes entre deux espèces qui n'avaient pas de plus proche voisin dans leur propre génome.

Enfin de nouvelles méthodes, dans leur esprit, sont apparues. Le "quartet puzzling" qui consiste à associer, pour tout quadruplet d'espèces, les deux paires qui s'opposent. Puis, ceci étant fait, d'assembler tous ces sous-arbres à quatre feuilles en en un arbre unique compatible avec un nombre maximum d'ente eux. La méthode BioNj, basée sur un principe d'évolution minimum (2002) a rapidement balayé le standard NJ. Puis vint la domination des méthodes de maximum de vraisemblance, très sensiblement améliorées au LIRMM où les auteurs ont prouvé leur efficacité (2003).

De même les mesures de distance entre gènes ont rapidement évolué, en voulant s'affranchir des alignements. D'abord avec les fréquences comparées de courts nucléotides de longueur fixée au long des séquences (signature) ou de fragments identiques et maximaux de longueur quelconque mais uniques retrouvés dans un ordre quelconque (Maximum Uniq Matches, 2003). Mais tout ceci n'avait de sens que pour des séquences relativement proches. De même les méthodes basées sur l'ordre des gènes (Reversal distance) qui n'ont dû leur succès parmi les méthodologues que pour leur efficacité, un algorithme linaire ayant été établi.

Plus utiles ont été les méthodes de reconstruction basées sur des distances partielles, certaines valeurs entre paires d'espèces n'ayant pu être évaluées (faute de gènes orthologues). De même, de nombreuses méthodes de comparaison des topologies d'arbres ont été proposées, rapidement accompagnées de calcul d'un arbre consensus fait des sous-arbres majoritaires (2003) ou du plus grand sous arbre commun. Plus utiles encore ont été les progrès dans la représentation des arbres et même des représentations conjointes de deux arbres ayant les mêmes feuilles. Les spécialistes d'algorithmique combinatoire s'en sont donnés à coeur joie !

Enfin je terminerai cette énumération des questions traitées à Alphy par deux importants problèmes, celui des réordonnancement des gènes entre les chromosomes de deux génomes, inauguré par une équipe de l'ENS avec un petit poisson et l'homme et celui des co-évolutions dans lequel on observe les mutations simultanées de deux espèces vivant en symbiose.

Depuis mon éloignement de la communauté, d'autres méthodologies sont certainement apparues et je compte sur votre aide pour compléter cette liste.