Assemblage ciblé : recherche d'une famille de gènes sur un génome non assemblé

Mathieu Giraud[1] Pascale Quignon[2] Élodie Retout[1] Emmanuelle Morin[1]
Anne-Sophie Valin[1] Dominique Lavenier[1] Maud Rimbault[2] Francis Galibert[2] Jacques Nicolas[1]

[1] IRISA / INRIA / CNRS / Université de Rennes 1, 35 042 Rennes Cedex
jacques.nicolas@irisa.fr
[2] UMR 6061 / CNRS / Université de Rennes 1
galibert@univ-rennes1.fr

(Version mise à jour le 15 juillet 2005)

Résumé *L'assemblage ciblé est une méthode pour identifier rapidement une famille de gènes à partir d'un ensemble réduit de traces de séquençage. Cet ensemble est créé sur la base de la présence de motifs caractéristiques de la famille. Cette méthode court-circuite la période, souvent longue, qui sépare le séquençage de l'assemblage complet d'un génome. Elle permet également de contrôler finement les paramètres d'assemblage de la famille de gènes considérée. Pour illustrer le potentiel de l'assemblage ciblé, nous avons identifié sur le séquençage $7.6\times$ non assemblé du chien plus de 400 nouveaux gènes de récepteurs olfactifs en plus des 633 précédemment répertoriés. Ces résultats ont été confrontés avec une première version de l'assemblage.*

Keywords: assemblage, découverte de gènes, recherche de motifs

1 Introduction

Aujourd'hui, la méthode de séquençage usuelle, largement automatisée, est celle dite par *shotgun*. Cette méthode consiste à découper un génome complet en fragments qui sont insérés dans des vecteurs de clonage [2,8,23]. Les vecteurs sont intégrés dans des cellules hôtes qui assurent le maintien et la replication des fragments d'ADN. L'ensemble de ces cellules constitue une *banque de clones*. Chaque clone, dont la taille varie de 1 à 30 kB, est séquencé par ses extrémités 5' et 3', ce qui fournit des *traces* de 500 à 1200 nucléotides. Généralement, on considère que les techniques de séquençage actuelles permettent d'obtenir des traces fiables pouvant aller jusqu'à 1000 nucléotides. Cette limitation est due notamment au pouvoir résolutif du gel ou du polymère, au temps de migration ainsi qu'à l'épuisement des produits pour les fragments de grande taille.

Ensuite, on doit réaliser un *assemblage* des traces. Les traces qui présentent une fenêtre de chevauchement suffisante avec une bonne similarité sont mises bout à bout pour idéalement aboutir à un *contig* unique correspondant à un génome ou un chromosome entier. Plusieurs méthodes performantes d'assemblage existent, notamment les méthodes *overlap-layout-consensus* ou celles de graphes [7,20].

En réalité, l'assemblage produit un certain nombre de contigs de quelques milliers à quelques millions de nucléotides. La *finition* d'un assemblage consiste à positionner correctement les contigs. Au final, le procédé d'assemblage sur des génomes de milliards de bases prend plusieurs mois. En partant du même séquençage, il peut y avoir plusieurs versions de l'assemblage qui s'améliorent au fur et à mesure.

La publication de la séquence assemblée d'un génome est toujours un événement important, attendu par la communauté scientifique. Elle permet par exemple d'identifier et de localiser précisément les gènes sur

les chromosomes. Mais si l'assemblage complet est indispensable pour établir cette cartographie, il n'est, par contre, pas forcément nécessaire pour découvrir de nouveaux gènes : l'information est implicitement contenue dans les seules traces du shotgun. Par conséquent, l'identification de nouveaux gènes peut débuter dès que l'on dispose d'un nombre suffisant de traces. C'est sur cette idée que repose notre méthode : l'identification de gènes d'une même famille peut commencer sans attendre la publication finale de la séquence complète du génome.

Cependant, notre méthode nécessite d'abord de caractétiser la famille de gènes. Elle suppose donc quelques connaisances préalables, et en particulier d'avoir à disposition un échantillon de gènes déjà identifiés. À partir de celui-ci, on peut inférer des propriétés locales supposées caractériser cette famille comme, par exemple, des régions hautement conservées. Ces *domaines* de quelques acides aminés, préservés au cours de l'évolution, peuvent être représentés par des motifs, comme dans la syntaxe PROSITE [14] de type *KLP-[IV]-x(1,2)-T*. Des programmes tels que PRATT [15] sont capables d'extraire automatiquement de tels motifs à partir d'un jeu de séquences.

Une famille de gènes est donc caractérisée par la présence d'un ou plusieurs motifs. Cette information est l'élément clé qui va nous permettre de nous focaliser sur un nombre réduit de traces à partir desquelles nous serons en mesure d'identifier les gènes qui nous intéressent. La recherche de motifs sélectionnant les traces peut se faire de manière approchée, que cela soit par des techniques de programmation dynamique [25] ou par des automates pondérés [10]. D'autres modèles, parmi lesquels les chaînes de Markov, sont utilisables [9,12,16].

La suite de cet article se décompose de la manière suivante : la partie 2 présente en détail la méthode de l'assemblage ciblé. La partie 3 illustre notre approche sur la recherche de gènes codant des récepteurs olfactifs sur les données brutes du séquençage $7.6\times$ du chien. Nous montrons, en particulier, comment cette méthode à permis d'identifier 400 nouveaux gènes sur les 639 précédemment répertoriés.

2 Assemblage ciblé

L'assemblage ciblé à pour but de n'assembler qu'un petit nombre de traces (issues du shotgun complet d'un génome) pour obtenir directement les séquences d'une famille de gènes. La sélection appropriée de ces traces s'effectue sur la base de *propriétés locales* qui se retrouvent à la fois dans les gènes et dans les traces. Ces propriétés, qui s'expriment sur quelques *dizaines ou centaines* de nucléotides peuvent être capturées de différentes manières :

- par des *motifs*, exacts ou approchés, un gène pouvant contenir un ou plusieurs motifs ;
- par l'expression de *modèles* à l'aide de languages, d'automates, ou de chaînes de Markov [10,12,24] ;
- par la recherche de *similitudes* que l'on repère grâce au calcul d'alignements locaux. C'est une généralisation de la recherche de motifs avec substitutions, insertions et délétions. L'alignement local se calcule grâce à la programmation dynamique [25], et de nombreuses heuristiques telles que Blast [1] rendent possible des recherches sur des grandes banques.

Ces techniques peuvent s'appliquer à la découverte de gènes, mais aussi à d'autres applications comme la recherche de petits ARNs non fonctionnels.

La méthode d'assemblage ciblée est représentée par la figure 1. Elle se décompose en 4 étapes principales :

- (1) la *formalisation des propriétés locales* caractéristiques de la famille de gènes. Ces propriétés peuvent être inférées automatiquement à partir d'un échantillon ou être déterminées à partir de connaissances préalables ;
- (2) la *sélection des traces* : parmi toutes les traces disponibles on choisit celles qui exhibent les propriétés locales précédemment définies ;

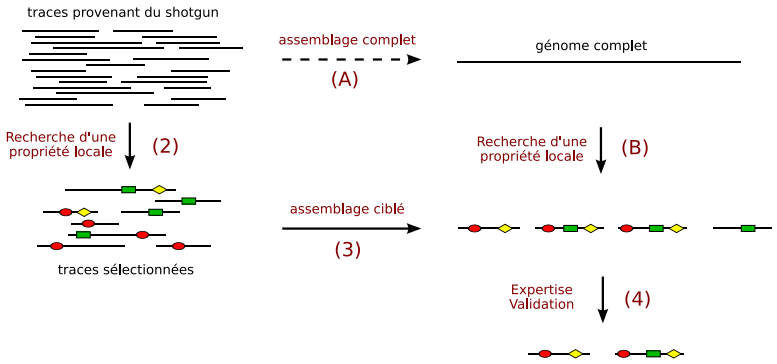

FIG. 1. L'assemblage complet d'un génome peut prendre plusieurs mois (A). Lorsque le génome complet est disponible, on utilise des techniques de découverte de gènes se basant sur des propriétés locales telles que la présence de motifs caractéristiques (B). L'assemblage ciblé consiste à filtrer directement les traces du séquençage (2) pour n'assembler qu'un petit sous-ensemble de ces traces (3). Une étape, non représentée, recherche de nouveau les propriétés locales sur les contigs produits par l'assemblage (3'). Dans les deux cas, des traitements supplémentaires avec une part d'expertise humaine sont nécessaires pour valider les gènes (4).

– (3) l'*assemblage* du sous-ensemble de traces retenues. À ce stade, on construit des contigs censés contenir les différents gènes de la famille ;
– (4) un *post-traitement* qui écarte certains contigs non pertinents ou construits par erreur.

Idéalement, les contigs produits par l'étape (3) correspondent exactement aux zones identifiées sur le génome complet par la recherche de motifs (B). Il faut pour cela que les traces sélectionnées à l'étape (2) soient suffisament nombreuses et chevauchantes.

Temps et qualité

On peut comparer les étapes d'un assemblage ciblé à celles des méthodes usuelles de découverte de gènes sur un assemblage complet :

– (1) la formalisation des propriétés locales, par exemple par des motifs, se fait de la même manière que lors d'une étude sur le génome assemblé, tout en veillant à ce que les propriétés locales puissent sélectionner suffisament de traces ;
– (2) La première recherche de motifs se fait sur des données plus volumineuses que celles du génome assemblé (1 à 15 fois) ;
– (3) l'assemblage se fait sur un ensemble beaucoup plus petit de traces, en fonction de la sélectivité de l'étape précédente (50 à 1000 fois plus petit) ;
– (3') une seconde recherche de motifs permet de préciser quels motifs apparaissent sur les contigs assemblés ;
– (4) enfin, le post-traitement, pouvant inclure une expertise humaine, est analogue à celui réalisé dans les méthodes usuelles.

L'assemblage ciblé est donc bénéfique dès que le temps gagné sur l'assemblage (3) est plus important que celui perdu sur les recherches de motifs (2+3'). En pratique, la recherche de motifs (2) s'effectue en temps *linéaire* par rapport à la taille totale des traces, et la seconde recherche (3') s'effectue sur une quantité de données négligeable. Au contraire, les algorithmes d'assemblage usuels [7,20] utilisent des comparaisons de

traces deux à deux et sont *quadratiques* dans la taille des traces. Le gain de temps va donc être important pour des sous-ensembles de traces beaucoup plus petits.

L'assemblage ciblé peut aussi permettre de choisir les paramètres d'assemblage plus adaptés à la famille de gènes étudiée qu'à un assemblage global. En particulier, si la famille est connue pour son fort taux de similarité, on peut fixer un seuil élevé d'identité pour l'assemblage afin de ne pas mélanger plusieurs gènes très similaires.

3 Recherche d'une famille de gènes de récepteurs olfactifs

Dans cette partie, nous appliquons l'assemblage ciblé à la recherche de gènes codant les récepteurs olfactifs dans le séquençage $7.6\times$ du chien. Cette recherche confirme l'intérêt de l'assemblage ciblé puisque 1083 gènes ont pu être identifiés sans utiliser le génome complet. Une discussion portant notamment sur la pertinence du choix des motifs conclut cette partie.

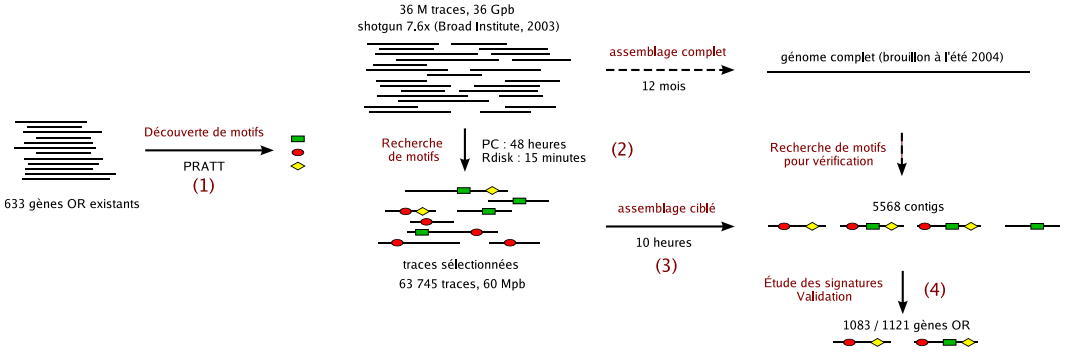

FIG. 2. Assemblage ciblé appliqué à la recherche de gènes OR chez le chien. À partir de 633 gènes précédemment connus, cinq motifs ont été découverts par PRATT (1), ce qui a sélectionné moins de $0,2\%$ des traces (2). Une fois assemblés (3), les contigs ont été étudiés (4) en utilisant notamment les signatures (figure 4). Le génome complet a permis de confirmer les contigs.

Les récepteurs olfactifs

Les récepteurs olfactifs (*olfactory receptors*, OR) ont été découverts par Buck et Axel en 1991 [3] chez le rat, découverte menant au prix Nobel de médecine 2004. Ils font partie de la famille des protéines transmembranaires couplées aux protéines G. Ils présentent sept domaines transmembranaires, certains étant relativement conservés, d'autres plus spécifiques au ligand. Chaque récepteur olfactif peut reconnaître plusieurs molécules odorantes, et une même molécule peut se lier à plusieurs récepteurs [17].

Les gènes codant les récepteurs olfactifs (appelés dans la suite « gènes OR ») ont une phase codante (ORF) d'environ 1000 pb. L'ORF est entièrement contenue dans un exon, ce qui facilite la recherche sur les traces. La famille des gènes OR est une des plus grandes familles de gènes chez les mammifères : 700 à 1000 gènes chez l'homme (dont 60% de pseudogènes), 1300 à 1600 chez la souris et chez le rat (dont 20% de pseudogènes). Les séquences des gènes OR sont assez similaires (de 30% à 98% d'identité), rendant possible des erreurs d'assemblage sur certaines portions. Des études sur le répertoire des ORs canins ont été menées dans [22] et [18].

Séquençage du génome du chien

Avec 38 paires d'autosomes plus X et Y, le génome du chien comprend 2,4 Gpb. Le premier séquençage du chien (shotgun $1.7\times$) a été publié en 2000 [19]. En 2003, un nouveaux séquençage a été réalisé au Broad Institute (shotgun $7.6\times$) [5]. Les traces (36 millions, 36 Gb) étaient disponibles dès 2003 mais la première version de l'assemblage ne le fut qu'à l'été 2004.

Avant notre étude, 633 gènes OR canins avaient été identifiés par séquençage à partir d'amorces dégénérées ainsi que par des comparaisons de séquences utilisant Blast à partir du séquençage $1.7\times$ [22]. Indépendamment, Olender et al avaient identifié 971 ORs, en partie par une méthode similaire à l'assemblage ciblé sur les traces du séquençage $1.3\times$ de Celera [18].

Assemblage ciblé des gènes OR canins

(1) Découverte de motifs PRATT [15], disponible sur les serveurs de OUEST-genopole®[3], a été utilisé pour extraire des motifs à partir des 633 gènes OR précédemment connus. Nous avons retenu cinq motifs présents sur plus de 80% des 633 gènes OR. Les motifs sélectionnés se répartissent sur toute la longueur des gènes OR (figure 3).

p1	TM II	P-M-Y-x-[FL]-L-x(2)-[FL]-[AMS]-x(2)-[DE]	1
p2	TM III	L-x(3)-M-x(0,1)-Y-x-[FLR]-[LY]-x(2)-[FILV]-[ACS]	0
p3	TM III	L-x(1,3)-M-x-[FILY]-D-R-x(2)-A-[IV]-[CS]-x-P-L-x-[HY]-x(3)-[ILM]	3
p4	TM VI	K-x-[FL]-[AGHNST]-T-C-x-[AS]-H-x(3)-[AIV]	1
p5	TM VII	N-P-[FILMV]-[IV]-Y-[AGST]-[AILMV]-[KR]-x(2)-[DEKQ]	1

FIG. 3. Motifs retenus pour les gènes OR du chien. Les motifs sont situés le long de différents domaines transmembranaires (TM). Le motif 2 était utilisé sous la forme MAYDRY dans [4]. La dernière colonne indique les seuils d'erreurs, fixés pour sélectioner moins de 25000 traces par motif.

(2) Recherche de motifs sur les traces Nous avons utilisé une recherche par automates pondérés pour localiser les cinq motifs avec erreurs parmi les 36 Gpb de traces (WAPAM, [10]). Cette recherche, qui dure 48 heures sur un PC conventionnel (2 GHz, 728 Mo RAM), peut être accélérée sur l'architecture Rdisk qui utilise des processeurs reconfigurables FPGA [11]. Au total 63745 traces ont été sélectionnées sur les 36 millions. Les fichiers qualités associés au séquençage [21] ont été pris en compte pour finalement ne sélectionner que 61321 traces, soit une sélectivité de $1,5\%$. Les traces ont une longueur moyenne de 658 bases.

(3) Assemblage L'assemblage, effectué par CAP3 [13] avec une fenêtre de 25 nucléotides et un taux de similarité de 97%, a duré 10 heures. 5568 contigs et 11839 singletons ont été obtenus. Les singletons correspondent à des traces de moyenne qualité qui n'ont pas pu s'assembler avec d'autres et n'ont pas été considérés dans la suite. Les contigs assemblés ont en moyenne une longueur de 1130 bases et une profondeur de 7 traces.

[3] `www.genouest.org`

(4) Post-traitement des résultats Pour étudier les 5568 contigs, nous avons localisé de nouveau les motifs, puis représenté leur succession par des *signatures* du type `p1 (171) p2 (27) p3 (339) p4 (141) p5`. Ces signatures indiquent la distance entre les extrémités terminales de chaque motif. L'analyse de la répartition des signatures sur l'ensemble des gènes OR a permis de mettre en évidence une distribution consensus (figure 4). Les contigs ont été aussi comparés par Blast aux gènes OR existants ainsi qu'entre eux-mêmes. Étant donné que l'assemblage était relativement strict pour ne pas commettre d'erreurs, certains contigs très similaires ont été réassemblés. Les ORFs ont été déterminées. Au total, 1083 gènes OR ont été identifiés dont 213 pseudo-gènes.

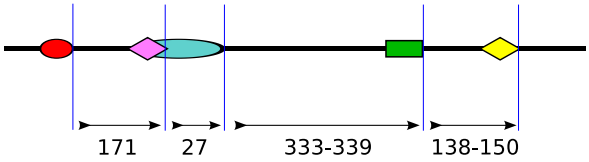

FIG. 4. Emplacement des motifs sur les gènes OR du chien. Les distances sont données en nucléotides. Cet emplacement correspond à la signature `p1 (171) p2 (27) p3 (333-339) p4 (138-150) p5`. Les motifs 2 et 3 sont imbriqués.

Validation des résultats

Lorsque le premier « brouillon » *(working draft)* du génome complet a été disponible (été 2004), la même recherche de motifs a été relancée sur l'assemblage complet [6]. Les séquences obtenues par l'assemblage ciblé ont été confirmées. L'assemblage complet a cependant permis d'étendre certains contigs lorsque l'ORF obtenue par l'assemblage ciblé n'était pas complète. De plus, certaines séquences provenant d'une étude sur le séquençage $1.7\times$ ont été ajoutées pour arriver à un répertoire total de 1121 gènes OR. Les contigs font en moyenne 1816 nucléotides, et les ORFs 950 nucléotides.

Parmi tous ces gènes, 125 soit 11% n'ont pas encore été localisés sur le génome (« chromosome inconnu ») alors que ces séquences représentent moins de 5% des séquences. En raison de la haute similarité de cette famille, il est possible que ces gènes soient difficiles à assembler et à cartographier.

Discussion

Pertinence du choix des motifs. L'approche utilisée par Olender dans [18] effectuait des recherches de traces avec BLAST/TBLASTN à partir d'un jeu de 199 séquences représentatifs des ORs connus chez les mammifères. Leur méthode recherche aussi des propriétés locales, à savoir des similarités d'au moins 40 paires de bases qui correspondent aux régions conservées dans les ORs. Une telle recherche augmente le temps de l'étape (2) de manière significative.

Nous avons fait ici le choix de partir d'une recherche de motifs pour se concentrer sur les zones similaires dans les ORs, ce qui autorise un seuil d'erreur élevé sur ces motifs. Certes, étant donné les seuils d'erreurs, la majorité des traces retenues à l'étape (2) ne se retrouvent plus dans les 1083 gènes OR finalement identifiés : il y a un grand nombre de singletons et la plupart des 5568 contigs formés sont des faux positifs. Le premier tri a d'ailleurs été d'analyser les contigs qui contenaient au moins le motif `p3`, motif le plus conservé dans les

gènes OR précédemment connus et utilisé dans certaines de ses formes dans [4]. Il apparaît que 94% des gènes OR finalement retenus possèdent ce motif, dont 72% sans erreurs.

Cependant, l'utilisation des 5 motifs avec erreurs a permis de sélectionner un sous-ensemble de traces plus important. On a ainsi plus de chances de trouver dans ce sous-ensemble suffisament de traces qui se chevauchent et s'assemblent pour élargir au maximum les contigs et idéalement trouver toutes les ORFs complètes. À titre de comparaison, l'utilisation d'un seul motif sans erreur, pourtant le plus pertinent (`p3`) conduit à seulement 862 contigs dont la séquence consensus a une longueur moyenne de 1083. Le motif `p3` étant très discriminant, de nombreux OR se trouvent dans ces 862 contigs, mais leur nombre reste inférieur aux 1083 gènes identifiés et les ORFs sont majoritairement incomplètes. De même, la méthode proposée dans [18] ne trouve que 121 ORFs complètes sur 971 ORs. Bien que, comparé à l'ensemble des traces du shotgun, l'assemblage ciblé traite très peu de traces, le sous-ensemble sélectionné doit contenir suffisament pour pouvoir être assemblé, quite à avoir un certain nombre de faux positifs à l'étape 2 qui sont par la suite éliminés.

Une perspective intéressante serait de trouver des modèles de propriétés locales plus complexes dans leur expressivité, mais qui restent relativement simples à calculer, comme par exemple des recherches par modèles de Markov cachés ou par automates pondérés.

Temps de calcul et expertise humaine. Outre le gain de temps non négligeable, l'assemblage ciblé permet de contrôler l'assemblage pour une famille donnée de gènes. Certes, le temps nécessaire à l'assemblage complet n'est pas seulement dû aux temps de calcul : la finition nécessite une expertise humaine et tire bénéfice d'expériences complémentaires, notamment sur la cartographie du génome.

L'avantage de l'assemblage ciblé est que le temps de calcul devient négligeable comparé au temps d'expertise. On peut aussi chercher à automatiser l'étape (4) a posteriori (en ne gardant que les signatures correctes), mais un travail d'analyse et de vérification sera toujours nécessaire. Plus généralement, dans une méthode d'assemblage ciblé, le travail humain peut se concentrer sur la famille de gènes étudiée.

La principale limitation de l'assemblage ciblé concerne la distance séparant les motifs sur les gènes. La présence de grands introns peut rendre impossible une sélection correcte des traces, sauf si ces introns présentent eux aussi une propriété locale caractéristique. Une autre limitation de l'assemblage ciblé est qu'il ne permet pas de localiser les gènes identifiés sur le génome, mais des amorces de PCR peuvent être dessinées afin de localiser ces gènes par une technique de cartographie génétique.

4 Conclusion et perspectives

L'assemblage ciblé est une méthode pour rechercher directement des propriétés locales dans un génome non assemblé. Lorsque des motifs communs à une famille de gènes sont connus, il permet de rechercher de nouveaux gènes directement sur les traces. L'assemblage ciblé a été validé par l'identification de nouveaux gènes olfactifs sur le séquençage $7.6\times$ non assemblé du chien. D'autres types de recherches sont envisageables tant que les propriétés recherchées sont suffisament locales et rapprochées.

Comme l'assemblage est réalisé sur un sous-ensemble restreint des traces, il est beaucoup plus rapide et ne nécessite pas une connaissance sur tout le génome : le temps d'assemblage se réduit de plusieurs mois à une journée. L'expertise humaine pour analyser les résultats peut donc s'appliquer directement sur un sous-ensemble pertinent du génome.

L'étape d'assemblage, adapté au problème considéré, gagnerait à être encore étudiée : au lieu d'utiliser un programme d'assemblage générique, nos recherches actuelles portent sur le développement d'algorithmes d'assemblages spécialement adaptés à des gènes hautement similaires.

Références

[1] S.F. Altschul, W. Gish, W. Miller, W.E. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3) :403–410, 1990.

[2] S. Anderson. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res.*, 10 :3015–3027, 1981.

[3] L. Buck and R. Axel. A novel multigene family may encode odorant receptors : a molecular basis for odor recognition. *Cell*, 65(1) :175–87, 1991.

[4] S. G. Conticello, Y Pilpel, G. Glusman, and M. Fainzilber M. Position-specific codon conservation in hypervariable gene families. *Trends Genet.*, 16(2) :57–59, 2000.

[5] Dog whole genome shotgun sequence traces. *ftp.ncbi.nih.gov/pub/TraceDB*.

[6] Dog genome sequence assembly Canfam 1.0. *www.genome.usc.edu*.

[7] B. Ewing and P. Green. Basecalling of automated sequencer traces using phred – error probabilities. *Genome Research*, 8 :186–194, 1998.

[8] R. C. Gardner and al. The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. *Nucleic Acids Res*, 9 :2871–2888, 1981.

[9] A. Gatiker, E. Gasteiger, and A. Bairoch. ScanProsite : reference implementation of a PROSITE scanning tool. *Applied Bioinformatics*, 1 :107–108, 2002.

[10] M. Giraud and D. Lavenier. Linear encoding scheme for weighted finite automata. In *Ninth International Conference on Implementation and Application of Automata (CIAA 2004)*, july 2004.

[11] S. Guyetant, M. Giraud, S. Derrien, L. Lhours, S. Rubini, F. Raimbault, and D. Lavenier. Cluster of reconfigurable nodes for scanning large genomic banks. *Parallel Computing*, 31(1) :73–96, 2005.

[12] D. Haussler, A. Krogh, I.S. Mian, and K. Sjolander. Protein modelling using hidden markov models : Analysis of globins. In *Hawaii Int. Conf. System Sciences*, January 1993.

[13] X. Huang and A. Madan. CAP3 : A DNA sequence assembly program. *Genome Research*, 9 :868–877, 1999.

[14] N. Hulo, C.J. Sigrist, V. Le Saux, P.S. Langendijk-Genevaux, L. Bordoli, A. Gattiker, E. De Castro, P. Bucher, and A. Bairoch. Recent improvements to the PROSITE database. *Nucl. Acids. Res.*, 32 :D134–D137, 2004.

[15] Inge Jonassen, John F. Collins, and Desmond Higgins. Finding flexible patterns in unaligned protein sequences. *Protein Science*, 4(8) :1587–1595, 1995.

[16] G. Kucherov and M. Rusinowitch. Matching a set of strings with variable length don't cares. *Theoretical Computer Science*, 178 :129–154, 1997.

[17] B. Malnic, J. Hirono, T. Sato, and L. Buck. Combinatorial receptor codes for odors. *Cell*, 96 :713–723, 1999.

[18] Tsviya Olender, Tania Fuchs, Chaim Linhart, Ron Shamir, Mark Adams, Francis Kalush, Miriam Khen, and Doron Lancet. The canine olfactory subgenome. *Genomics*, 83 :361–372, 2004.

[19] Elaine A. Ostrander, Kerstin Lindblad-Toh, and Eric S. Lander. Sequencing the genome of the domestic dog canis familiaris. www.genome.gov/11008069.

[20] Pavel A. Pevzner, Haixu Tang, and Michael S. Waterman. An eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci USA*, 98(17) :9748–9753, 2001.

[21] Quality files from NCBI. *www.ncbi.nlm.nih.gov/Traces/trace.cgi*.

[22] P. Quignon, E. Kirkness, E. Cadieu, N. Touleimat, R. Guyon, C. Renier, C. Hitte, C. André, C. Fraser, and F. Galibert. Comparison of the canine and human olfactory receptor gene repertoires. *Genome Biology*, 4 :R80, 2003.

[23] F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, and G. B. Petersen. Nucleotide sequence of the bacteriophage lambda DNA. *J. Mol. Biol*, 162 :729–773, 1982.

[24] D.B. Searls. The computational linguistics of biological sequences. In Larry Hunter, editor, *Artificial Intelligence and Molecular Biology*, pages 47–120. AAAI Press, 1993.

[25] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol*, 147 :195–197, 1981.