SEQAID II
UN LOGICIEL DE RECHERCHE POUR L'ENSEIGNEMENT DE LA GENETIQUE MOLECULAIRE

D. LENNE, J-F. RODES, N. SALAMÉ

SEQAID II est un logiciel de recherche qui comporte un ensemble de traitements sur des séquences nucléiques ou protéiques : visualisation, édition, comparaison, analyse, prédiction. Il fonctionne sur tous les compatibles PC. Accompagné d'une banque de données et d'une documentation détaillée, il est diffusé gratuitement dans tous les lycées dans le cadre d'une coopération entre l'INRP et la Direction des Lycées et Collèges.

1. L'INFORMATIQUE EN BIOLOGIE MOLÉCULAIRE

La biologie moderne s'appuie fortement sur les séquences de gènes et de protéines qui sont archivées par les chercheurs depuis plus de vingt ans. L'utilisation des moyens informatiques en biologie moléculaire a suivi rapidement l'évolution des techniques de séquençage et l'explosion du nombre de séquences disponibles. Dès le début des années 80 des bases de données internationales ont été mises en place aux Etats Unis, en Europe et au Japon. Les objectifs de ces bases sont de trois ordres : constituer des archives fiables, en permettre la consultation et la manipulation par les chercheurs en fonction de différents critères, y associer des outils d'analyse appropriés. Implantées sur de gros systèmes informatiques, ces bases sont accessibles, d'une part, à travers les réseaux de télécommunication, et d'autre part, au moyen de mises à jour régulières sur CD-ROM.

Il est exclu d'utiliser pour l'enseignement secondaire les logiciels qui fonctionnent sur ces systèmes puissants, et dont beaucoup sont des outils spécialisés pour réaliser à distance la recherche de séquences, leur extraction, la conversion des formats, les traitements statistiques, etc. Ces programmes regroupent souvent des modules et des algorithmes mis au point par différents chercheurs un peu partout dans le monde. Ils ne

peuvent se prêter à un travail pédagogique au lycée parce que la plupart d'entre eux ne comportent pratiquement pas d'interface et ne sont vraiment utilisables que par les chercheurs. En revanche, un nombre croissant de ces logiciels est porté sur les micro-ordinateurs en conservant très souvent les algorithmes implantés sur les gros systèmes.

Il existe ainsi de multiples programmes spécialisés qui réalisent, par exemple :

- la lecture de résultats d'électrophorèses ;
- la gestion des séquences déterminées localement ;
- la comparaison d'une séquence avec un ensemble d'autres séquences ;
- la recherche de signaux et de sites particuliers ;
- la recherche de régions codantes ;
- l'analyse des séquences (composition d'une protéine, codons utilisés dans une séquence nucléique, etc.) ;
- la prédiction de l'appartenance des acides aminés d'un polypeptide à des structures particulières (hélice, feuillet, coude, etc.) ;
- la construction de phylogénies à partir des différences entre séquences ;
- la visualisation de molécules dans l'espace.

Considérant les contenus des programmes actuels de biologie au lycée, des domaines tels que l'étude des maladies génétiques, des parentés entre les espèces, des relations entre la structure d'une molécule et sa fonction, pourraient bénéficier d'une partie de ces traitements à condition de disposer des outils logiciels et des données moléculaires appropriées sur les matériels qui équipent les établissements scolaires.

Peu de logiciels sur micro-ordinateur proposent simultanément tous les traitements. Globalement, on peut considérer que les outils qui intéressent en priorité l'enseignement sont ceux qui regroupent les fonctions d'analyse et d'édition de séquences, ceux qui sont dédiés aux comparaisons de séquences et à la construction de phylogénies, et ceux qui sont destinés à la visualisation et à la modélisation graphique de molécules.

Le logiciel SEQAID II constitue un des logiciels "généralistes" qui présente des fonctions développées pour l'édition (conversion, modification, assemblage, etc.) et l'analyse de séquences. Conçu par deux chercheurs à l'Université du Kansas, il est mis à la disposition de la

communauté scientifique sur le serveur européen de biologie moléculaire. Avec l'autorisation des auteurs, on a procédé à la réalisation d'une version française qui conserve les caractéristiques de l'outil de recherche tout en l'adaptant dans les détails à une utilisation pédagogique. Cette version présentée ici est maintenant à la disposition de tous les établissements scolaires. Par rapport à la version anglaise, une attention particulière a été accordée à la simplification de la mise en oeuvre, à la terminologie et à l'aide en ligne destinée aux élèves.

2. FONCTIONNALITÉS DU LOGICIEL SEQAID II

SEQAID II propose un nombre important de fonctionnalités concernant l'édition, la comparaison et l'analyse de séquences nucléiques et protéiques, mais il reste néanmoins peu volumineux et assez rapide. Sur un ordinateur de 640 Ko de mémoire, par exemple, le programme, écrit en Turbo Pascal, nécessite 300 Ko environ, le reste de la mémoire disponible pouvant accueillir les séquences nucléiques ou protéiques. Toutes les fonctions sont documentées, l'aide en ligne étant conçue pour les élèves seulement. Sauf dans les cas où la décision de l'utilisateur est indispensable, les fonctions s'exécutent avec des valeurs par défaut, ce qui accélère et simplifie les processus de traitement.

2.1 Modification de la configuration à l'installation

Le logiciel fonctionne, par défaut, avec un ensemble de fichiers (code génétique, données sur les acides aminés, configuration d'imprimantes, listes d'enzymes de restriction, de sites de régulation, etc.). L'option **Fichiers de configuration** permet de modifier, pour la session de travail seulement, le code génétique utilisé, le codage des acides aminés et leurs propriétés physiques (hydrophobicité, probabilité d'appartenance à des structures, masse molaire, etc.) ; on peut en outre changer la configuration de l'écran (couleur, monochrome) et la configuration de l'impression (deux fichiers de configuration sont fournis, l'un pour une imprimante à aiguilles, l'autre pour une imprimante laser, mais d'autres paramètres adaptés à d'autres imprimantes, peuvent être ajoutés). Les fichiers des sites de restriction (plus d'une centaine) et de régulation (une douzaine) fournis également avec le logiciel, sont des fichiers texte standard. Ils peuvent être modifiés et enrichis avec n'importe quel éditeur, à condition de respecter le format indiqué en tête de chaque fichier.

2.2 Lecture, écriture, visualisation de séquences

Le logiciel travaille uniquement sur des séquences contenues en mémoire ; une trentaine peuvent être présentes en mémoire simultanément, pour un volume global de 300 Ko environ. Diverses fonctions du logiciel permettent d'accéder à n'importe quel répertoire existant (pour charger, sauvegarder, voir le contenu, renommer un fichier, parcourir un répertoire, modifier le chemin d'accès à un répertoire de travail, etc.). Le logiciel est ainsi totalement ouvert sur l'environnement. Il convient de noter que SEQAID II a son propre format de données dans lequel doivent être mises les séquences importées, ce qui peut être fait aisément avec un éditeur de texte. Chaque séquence doit être contenue dans un fichier particulier. La séquence proprement dite peut être précédée d'un nombre variable de lignes de commentaires. Il y a très peu de modifications à introduire pour lire les formats des banques de données internationales (qui diffèrent seulement par les lignes de commentaires).

Plusieurs options permettent de préparer de manière très souple l'affichage à l'écran, l'impression sur papier et l'exportation de séquences ou de résultats dans des fichiers. Le logiciel gère lui-même l'impression sur papier de façon à obtenir des documents adaptés à une publication scientifique ou à des élèves : jusqu'à 240 caractères par ligne, en mode condensé ou étendu, en double ou simple frappe, en double ou simple interligne, avec ou sans marge, etc.

2.3 Edition de séquences

Ces options permettent à l'utilisateur de mettre en forme ses données et d'enrichir sa banque par l'entrée de séquences au clavier (en fonction de la séquence à créer - nucléique ou protéique - des filtres sont utilisés pour contrôler les caractères frappés et éviter les erreurs) ou aussi par la fabrication de nouvelles séquences à partir d'autres séquences archivées : on peut procéder par extraction de parties de séquences, par concaténation, ou par troncature.

2.4 Conversion de séquences

Les conversions concernent des opérations à visées éditoriales (duplication, changement de casse, inversion, etc.) et des opérations qui ont une signification biologique précise telles que :

- le passage de l'ADN à l'ARN : en partant du brin non transcrit du gène et en remplaçant simplement les T par des U, on obtient l'ARNm correct mais sans passer par le processus de la transcription ; il s'agit donc bien d'une simple conversion ;

- le passage d'un ADN ou d'un ARN à un polypeptide : le point de départ de la "traduction" est précisé par l'utilisateur, le programme s'arrêtant au premier codon stop rencontré. A la suite de cette "traduction", on peut visualiser la protéine seule ou alignée avec la séquence nucléique, examiner les différents codons utilisés dans la séquence nucléique ou la composition de la protéine en acides aminés. La séquence résultant d'une traduction peut être conservée en mémoire puis sauvegardée sur disque ;
- l'affichage des différentes séquences nucléiques possibles étant donné le caractère dégénéré du code génétique ;

Séquences nucléiques possibles de ALPHA.PRO

Les bases de l'ADN sont A C G T (N=toute base, R=purine, Y=pyrimidine)

```
Met Val Leu Ser Pro Ala Asp Lys Thr Asn Val Lys Ala   13
ATG GTN TTR TCN CCN GCN GAY AAR ACN AAY GTN AAR GCN
        CTN AGY

Ala Trp Gly Lys Val Gly Ala His Ala Gly Glu Tyr Gly   26
GCN TGG GGN AAR GTN GGN GCN CAY GCN GGN GAR TAY GGN

Ala Glu Ala Leu Glu Arg Met Phe Leu Ser Phe Pro Thr   39
GCN GAR GCN TTR GAR CGN ATG TTY TTR TCN TTY CCN ACN
           CTN      AGR          CTN AGY

Thr Lys Thr Tyr Phe Pro His Phe Asp Leu Ser His Gly   52
ACN AAR ACN TAY TTY CCN CAY TTY GAY TTR TCN CAY GGN
                                  CTN AGY

Ser Ala Gln Val Lys Gly His Gly Lys Lys Val Ala Asp   65
TCN GCN CAR GTN AAR GGN CAY GGN AAR AAR GTN GCN GAY
AGY

Ala Leu Thr Asn Ala Val Ala His Val Asp Asp Met Pro   78
GCN TTR ACN AAY GCN GTN GCN CAY GTN GAY GAY ATG CCN
    CTN

Asn Ala Leu Ser Ala Leu Ser Asp Leu His Ala His Lys   91
AAY GCN TTR TCN GCN TTR TCN GAY TTR CAY GCN CAY AAR
        CTN AGY      CTN AGY      CTN

Leu Arg Val Asp Pro Val Asn Phe Lys Leu Leu Ser His  104
TTR CGN GTN GAY CCN GTN AAY TTY AAR TTR TTR TCN CAY
CTN AGR                           CTN CTN AGY

Cys Leu Leu Val Thr Leu Ala Ala His Leu Pro Ala Glu  117
TGY TTR TTR GTN ACN TTR GCN GCN CAY TTR CCN GCN GAR
    CTN CTN          CTN               CTN

Phe Thr Pro Ala Val His Ala Ser Leu Asp Lys Phe Leu  130
TTY ACN CCN GCN GTN CAY GCN TCN TTR GAY AAR TTY TTR
                        AGY CTN               CTN

Ala Ser Val Ser Thr Val Leu Thr Ser Lys Tyr Arg      142
GCN TCN GTN TCN ACN GTN TTR ACN TCN AAR TAY CGN
    AGY     AGY          CTN      AGY          AGR
```

- la détection du segment d'un séquence nucléique dont la traduction est susceptible de fournir un polypeptide donné (segment commençant par un codon d'initiation et se terminant par un codon stop) ;

```
SEQAID II- Recherche de segments dans GAMMADNC.ADN

Liste des segments de    441  bases

  Début       Fin      Cadre     Acides aminés
  -----      -----     -----     -------------
    54         495       3           147
```

Le logiciel a trouvé dans le 3ème cadre de lecture un segment qui commence en 54, se termine en 495, et qui code pour un polypeptide de 147 acides aminés.

- la recherche, dans une séquence nucléique, de tous les segments susceptibles d'être traduits en acides aminés, dans tous les cadres de lecture (que la traduction commence à partir de la 1ère, la 2ème ou la 3ème base).

2.5 Les comparaisons de séquences

Trois types de comparaisons sont possibles :

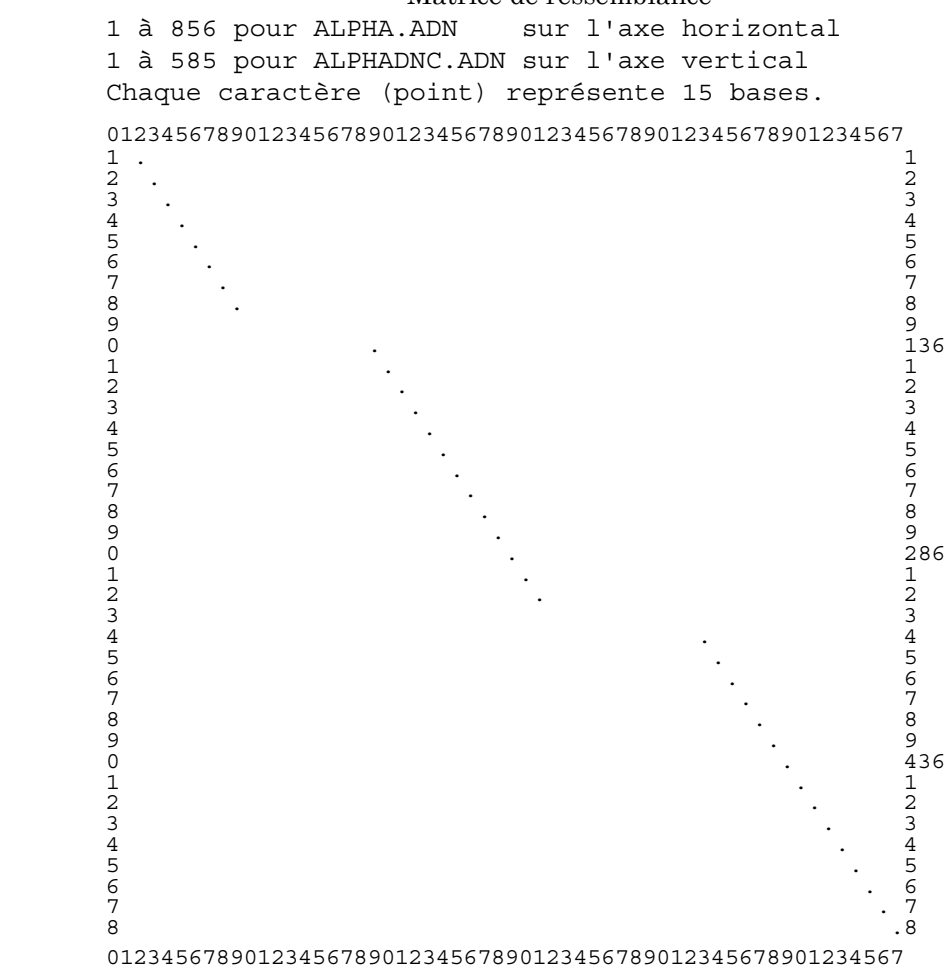
- comparaison simple par la mise en parallèle de deux séquences : les positions de départ peuvent être recherchées automatiquement par le logiciel ou bien fixées par l'utilisateur. A l'affichage, les éléments différents sont matérialisés par un code couleur ; les deux séquences peuvent ensuite être parcourues dans les deux sens et éventuellement modifiées.

```
1  dans BETACOD.ADN                        BETACOD.ADN     60
v        v         v         v         v         v         v
ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAAC
ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGAAGGTGAACG
v        v         v         v         v         v         v
1  dans THA6COD.ADN                        THA6COD.ADN     60
```

Affichage des 60 premiers éléments mis en parallèle. Les deux séquences sont identiques jusqu'à la 50ème base. Ici, les bases identiques sont en caractère normal et les bases différentes sont en caractère gras. Sur l'écran, les différences sont en rouge. Les signes v et ^ sont des taquets toutes les dix positions.

- recherche d'une similitude globale entre deux séquences : le programme construit un graphique avec l'une des séquences en x et l'autre en y. Les portions identiques montrent un alignement de points sur la diagonale ;

Matrice de ressemblance



Il y a d'abord une partie qui existe dans la première séquence seulement, puis une partie commune, suivie d'une partie présente dans la première séquence, etc.

- alignement avec recherche de discontinuités : cette comparaison est adaptée lorsque deux séquences peuvent avoir en partage

plusieurs segments identiques, séparés par des régions différentes. Le logiciel recherche les meilleurs alignements possibles (qui traduisent plus ou moins bien la similitude entre les deux séquences) ;

- on peut associer à ces comparaisons une fonction qui calcule, s'agissant de séquences nucléiques, la probabilité pour qu'un nombre donné d'éléments identiques soit dû au hasard. Comme on dispose de quatre bases, il faut que la ressemblance entre deux séquences concerne plus de 25 % de leurs éléments pour qu'on puisse chercher l'explication de la ressemblance par un facteur autre que le hasard.

2.6 Analyse de séquences

On peut regrouper dans cet ensemble les fonctions suivantes :

- description de la composition d'une séquence nucléique (décompte des codons utilisés) ou protéique (acides aminés utilisés et calcul de la masse molaire de la protéine) ;
- calcul de la taille et de la mobilité de fragments d'ADN sur un gel d'électrophorèse ;
- recherche de sites particuliers dans l'ADN : cette fonction permet de rechercher une suite quelconque d'éléments, donc l'établissement d'une carte de restriction d'une séquence ;
- la localisation dans une séquence anonyme de parties probablement codantes, en utilisant les fréquences des codons utilisés dans des séquences codantes connues ;
- la prédiction de l'hydrophobicité d'une protéine et de son antigénicité, le calcul de la probabilité pour que des parties aient une structure secondaire du type hélice alpha ou feuillet bêta, etc.

3. DONNÉES ASSOCIÉES ET UTILISATIONS PÉDAGOGIQUES

SEQAID II est diffusé accompagné d'une banque de séquences sur les globines (séquences des chaînes alpha, bêta, gamma et delta de l'hémoglobine humaine, protéines correspondant à ces chaînes, allèles du gène bêta de l'hémoglobine responsable de différentes hémoglobinopathies) et le système des groupes sanguins ABO.

On a limité intentionnellement les données distribuées dans un premier temps afin de laisser les utilisateurs prendre en mains le logiciel sur des exemples bien connus. Les séquences sont structurées par section

en fonction des connaissances exigibles : il y a donc des répertoires et des données spécifiques pour les classes de première littéraire, économique et sociale, scientifique, ainsi que pour les classes de terminale. On peut aborder avec ces données, dans le cadre du programme commun :

- l'étude de l'expression d'un gène déterminé ;
- les relations génotype-phénotype à partir de la comparaison des allèles d'un gène et de celle des polypeptides qu'ils codent ;
- la notion de famille multigénique.

Des données complémentaires sont déjà constituées pour une diffusion à court terme : système HLA (gènes d'histocompatibilité de classe I), gène responsable de la "phénylcétonurie", hormones, séquence du gène du récepteur aux LDL. Ces données, auxquelles s'ajoute un autre ensemble en cours de constitution, seront regroupées autour de thèmes :

- facteurs génétiques et nutritionnels dans l'expression de certaines maladies ;
- polymorphisme génétique et médecine prédictive ;
- génétique humaine associant l'analyse d'arbres généalogiques à celle des données moléculaires sur les allèles transmis ;
- complexification du génome au cours de l'évolution.

4. CONDITIONS DE DIFFUSION

Un exemplaire de SEQAID II est diffusé gratuitement dans tous les lycées. Les utilisateurs sont autorisés à copier programmes et données, et à reproduire les documents issus du livret d'accompagnement pour les élèves. En revanche, ils n'ont pas le droit de supprimer les fichiers, de les altérer, ou d'en faire un usage marchand. Les documents d'accompagnement comportent une prise en mains guidée à l'intention des enseignants, des exemples de mise en oeuvre pédagogique, et une documentation détaillée sur les fonctions du logiciel. Un questionnaire est annexé pour recueillir les appréciations des enseignants sur l'ensemble logiciel - données - documentation. Les réponses à ce questionnaire sont destinées à évaluer dans quelle mesure les outils proposés correspondent aux besoins, et à préparer, s'il le faut, une nouvelle version du logiciel.

Le suivi de cette opération est réalisé par l'intermédiaire d'un service télématique mis en place sur le serveur de l'INRP. Appeler le 3616, code INRP (tarifs février 1994 : 0,12 F puis 0,99 F la minute).

RÉFÉRENCES

- BELL G. I., MARR T. (eds) : *Computers and DNA*. Addison-Wesley, New York, 1989.
- BISHOP M. J., RAWLINGS C. J. (eds) : *Nucleic acid and protein sequence analysis*. A practical approach, IRL Press, 1987.
- COLWELL R. R. (ed) : *Biomolecular Data : A Resource in Transition*. Oxford University Press, Oxford, 1989.
- DOOLITTLE R.F. (ed.) : *Molecular Evolution : Computer Analysis of Protein and Nucleic Acid Sequences*. Methods in Enzymology, Vol 183, Academic Press Inc., San Diego, 1990.
- DOOLITTLE R.F. (ed.) : *Of URFS and ORFS. A primer on how to analyse derived amino acid sequences*. Oxford University Press, 1986.
- GRIBSKON M., DEVEREUX J. : *Sequence analysis primer*. Macmillan, Stockton Press, 1991.
- HEIJINE G. VON : *Sequence analysis in molecular biology. Treasure trove or trivial pursuit*. Academic Press Inc, 1987.
- LESK A.M. (ed.) : *Computational Molecular Biology. Sources and Methods for Sequence Analysis*. Oxford University Press, Oxford, 1988.
- LESK A.M. : *Protein architecture. A practical approach*. IRL Press, 91.

SEQAID II : Auteurs D.D. ROUFFA et D.J. RHOADS, Biology Division, Centre of Basic Cancer Research, Ackert Hall, Kansas State University, Manhattan, KS 66506, USA.

Le document d'accompagnement du logiciel est : HERVE J.C., SALAME N, THERRIE B. : Analyse de séquences de gènes et de protéines avec le logiciel SEQAID II. INRP, 1993.

Des exemplaires sont disponibles à l'Institut National de Recherche Pédagogique, Service des publications, 29, rue d'Ulm, 75230 - PARIS CEDEX 05 (Prix 100F. TTC ; chèque : M. l'agent comptable de l'INRP).

D. LENNE, J-F. RODES, N. SALAMÉ (INRP)