



Stages 2011-2012

Contact : Hervé PERDRIX, stage@isoft.fr

ISOFT : UNE SOCIETE INNOVANTE

ISoft est un **éditeur de logiciels** spécialisé dans le **traitement, l'exploration et l'analyse interactive de très gros volumes de données** et plus particulièrement dans la modélisation des traitements et des connaissances, l'intelligence artificielle et l'analyse de données. Ses produits principaux sont :

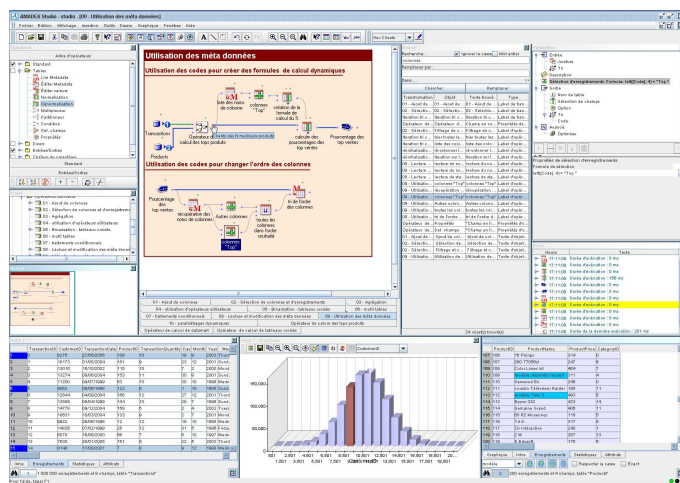
- **Alice**, logiciel de **Data-mining**,
- **Amadea** : plateforme d'analyse et de **traitement de données** temps réel basée sur le **Data-Morphing**,
- **AmadeaBiopack** : extension d'Amadea dédiée à l'analyse des données biologiques.
- **StatMining** ; logiciel de construction de modèles statistiques

ISoft a pour vocation d'apporter sur le marché des **solutions innovantes** mettant en œuvre de nouvelles technologies. Pour ce faire, ISoft a une activité de **Recherche** importante. Société technologique en croissance, indépendante, exclusivement tournée vers **les nouvelles technologies** et disposant d'un environnement technique de premier plan, ISoft propose des **stages** et des **opportunités de carrière** au sein de son département bio-informatique:

Les stages, proposés tout au long de l'année et généralement d'une durée minimale de 4 mois, sont destinés en priorité à des BAC + 5 ayant à la fois des connaissances en biologie et en informatique et qui désirent évoluer dans le domaine des nouvelles technologies. Les candidats ont un goût prononcé pour l'innovation et souhaitent évoluer au sein d'une **équipe de professionnels expérimentés** et d'un **environnement technique de qualité**.

La **plate-forme d'analyse de données biologiques Amadea Biopack**, permet à un biologiste d'explorer et analyser sans programmation ses résultats expérimentaux, quel que soit leur volume et leur nature. Grâce à une interface interactive, il enchaîne graphiquement des blocs fonctionnels permettant, en temps réel, de **relier ses données aux bases de données biologiques**, et ainsi d'en **extraire l'information recherchée**. Cette approche de la bio-informatique permet au bio-analyste d'éviter de recourir systématiquement à des scripts ou à la programmation d'outils informatiques pour répondre à des questions biologiques, et donc de **se focaliser sur sa recherche**.

Vous trouverez ci-dessous nos offres de stages en Bio-Informatique.





STAGES RECHERCHE

Les stages proposés ci-après ont pour but de participer à l'analyse de données biologiques à haut débit, en créant des pipelines intégratifs, enchaînant de manière originale, sources de données biologiques de tous types et outils existants. Les stages commencent généralement par une étude de la bibliographie et de l'état de l'art et débouchent sur des prototypes permettant de démontrer et valider l'approche retenue.

Les stagiaires sont intégrés aux équipes de Recherche et bénéficient des matériels et logiciels les plus récents.

Des connaissances préalables sur les sujets abordés lors des stages ou des techniques comparables seront appréciées.

Stage : " Analyse génomique de l'évolution d'une clade de levure "

Dans la cadre d'un projet collaboratif avec 4 laboratoires de recherche français de premier plan, le stage consiste à participer à une étude intégrative de l'évolution des génomes de nombreuses espèces de levures. Il s'agit d'une première approche multidisciplinaire et globale au niveau des génomes complets de tout un arbre de l'évolution d'une espèce. Il s'agira de mettre en place des processus originaux d'analyse des données issues des dernières machines de séquençage qui viendront supporter les expérimentations biologiques.

Stage : " Outils d'analyse des résultats de séquençage haut débit "

Les nouvelles technologies de **séquençage** à haut débit génèrent aujourd'hui des volumes de données tels que les outils existants ne savent plus les analyser, et que la recherche en biologie est actuellement dans un **goulot d'étranglement** dû à l'absence de méthodes efficaces pour les analyser. Or, la plateforme bio-informatique d'ISoft permet d'explorer tout type de données, quel que soit son volume, de manière interactive et sans programmation, ce qui permet d'envisager la mise au point aisée des outils manquant pour l'analyse des données de séquençage à haut débit.

Le stage consiste à analyser la structure des différents types de résultats obtenus par les **séquenceurs** à haut débit, et à créer les modules nécessaires dans la plateforme bio-informatique d'ISoft pour se connecter à tous ces résultats, et les analyser.

Stage " Bibliothèque pour le génotypage à très haute résolution "

Les puces **génotypage**, basées sur l'évaluation des polymorphismes à des positions données, permettent une analyse combinée, à **très haute résolution**, du nombre de copies et de l'équilibre allélique. De nombreux outils existent pour réaliser ce type d'analyse, essentiellement sous forme de scripts écrits en différents langages : R, Perl, Python ... Les chercheurs qui souhaitent utiliser ces outils doivent donc obligatoirement avoir une connaissance assez poussée des langages de programmation.

Le but du stage est donc de développer une bibliothèque d'outils dédiés à l'utilisateur final pour analyser ce type de données. Ces outils, **plus conviviaux et accessibles** à la majorité des chercheurs, pourront soit être spécifiquement créés, soit encapsuler les scripts existants.

Stage " Mise à jour automatique des informations biologiques "

La recherche en bio-informatique fait usage d'un nombre important de bases de **données de référence**, interdépendantes et mises à jour régulièrement par la communauté. Disposer de versions à jour et **cohérentes** entre elles est un défi important en raison des formats hétérogènes et des volumes mis en jeu. Le stage consistera à étudier **l'évolution des bases** importantes (NCBI, EMBL, UCSC, KEGG, ...) et à proposer des solutions automatisées globales ou incrémentales de récupération des données et de leurs **mises à jour**. La solution retenue sera illustrée par la réalisation d'un prototype intégré à une plateforme bio-informatique.



Stage "Bio statistiques"

L'analyse des **données biologiques** fait appel à des méthodes statistiques connues et éprouvées, et implémentées par de nombreux outils open-source. Toutefois, les **volumes** de données à traiter font apparaître les limites rédhibitoires de ces outils en terme de performances. Le stage consistera à tester des fonctions statistiques génériques et éventuellement les **adapter** pour les rendre compatibles avec les contraintes de volume liées aux données biologiques. Ce stage s'adresse à un étudiant ayant de bonnes bases en statistiques et pourra nécessiter des développements en C++.

Stage " Fonctions de calculs parallélisés et distribués "

La généralisation des environnements multi-cœurs et des réseaux hauts-débits permet d'envisager des applications de **traitement de données biologiques distribués**. Cette distribution est rendue nécessaire pour compenser l'augmentation exponentielle du volume des données à traiter. Le stage consistera à **rendre distribuable** des programmes qui sont aujourd'hui localisés sur un serveur. Le stage commencera par l'étude de différentes options de distribution de traitements et des données. Le stagiaire implémentera ensuite les couches logicielles nécessaires pour gérer **l'exécution à distance** de processus de traitement de données déportés (rsh, ssh, web services). Le stage pourra donner lieu à des développements en C++.

Stage " Croisement « intelligent » de données biologiques "

Les sources de données biologiques sont partielles car spécialisées, et la plupart du temps nécessitent d'être **croisées** avec d'autres sources pour en **extraire** le maximum **d'informations** utiles. Par exemple, étant donnée une liste d'identifiants d'objets biologiques (gènes, protéines), on cherchera à retrouver automatiquement les informations (gènes, exons, introns, fonction, pathway, ...) connues sur ces objets. Le stage consistera d'abord à **identifier** les méta données à associer aux différentes sources de données (NCBI, EMBL, UCSC, KEGG, ...) pour les qualifier, avant de les utiliser pour proposer des mécanismes transparents permettant d'inférer la manière d'obtenir une information à partir d'une autre.

Stage " Serveur d'application biologique "

Isoft dispose d'une **plateforme bio-informatique** permettant de définir des **pipelines d'analyse** et de les rendre disponibles sur le web. Le stage consistera à utiliser cette plateforme bio-informatique pour développer un site de démonstration des capacités de la plateforme. Le stagiaire choisira dans un premier temps les applications à **mettre en valeur** et les mettra à disposition sur le site après les avoir paramétrées. Dans un second temps, de nouvelles applications **génériques** destinées à rendre possible **l'exploration** de leurs données par les biologistes seront développées et mises à disposition.

The image displays several screenshots from the AmaMineReg software interface. On the left, a table lists genomic features with columns for Name, Chromosome, Start, End, Strand, and various scores. Below this is a line graph titled 'Motifs found inside promoter or gene: frequency relative to distance to a gene', showing peaks for motifs 1 through 5. In the center, a 'Gene Region' diagram shows a gene structure with 'Promoter', 'Inside', and 'Intergenic' regions. On the right, a complex pipeline diagram shows the workflow from 'Input' to 'Output' through various processing steps like 'Filtering', 'Clustering', and 'Annotation'. The bottom right shows a detailed view of a specific motif or gene region with associated data and visualizations.