GenePhys
(Projet de Recherche 2004)

ACI IMPBIO

Algorithmes efficaces et lois physiques réalistes pour les modèles de structures et d'analyses de séquences à grandes échelles. Applications des modèles structuraux aux analyses des génomes.

www://genephys.pasteur.fr

CONTEXTE

L'implémentation de modèles réalistes en bioinformatique se heurte souvent à l'obstacle méthodologique majeur de complexités algorithmiques impratiquables. Ces complexités sont le plus souvent rattachées aux effets à longue portée, qui sont par ailleurs indispensables pour le réalisme des modèles (les effets à longue portée amenant en 'contact' des éléments distants sur les séquences primaires). Dans la plupart des cas ce constat conduit à l'adoption de lois très simplifiées, que ce soit pour les modèles structuraux (représentation des boucles multiples dans les ARN, par exemple), ou pour les modèles d'analyses de séquences de la bioinformatique (représentation des 'gaps' dans les alignements, par exemple). Les enjeux pour surmonter cette situation d'impasse méthodologique sont nombreux, avec la perspective de modèles réalistes beaucoup plus précis pour les analyses de séquences et pour les modèles structuraux.

OBJECTIFS

Sur la base des travaux déjà accomplis par des membres du projet, il apparaît que pour une classe de modèles à valeur d'archétype il exsite une solution conceptuelle unique permettant de réduire de façon drastique les temps calcul (jusque des millions de fois), tout en tenant compte d'effets à longue portée réalistes . Cette méthode appellée SIMEX, développée au départ pour des modèles structuraux, repose sur la représentation numérique des effets à longue portée en sommes d'exponentielles. Le projet pouvait alors dans un contexte privilégié, se fixer les objectifs suivants :

-Rétablir des ponts entre les modèles biophysiques et bioinformatiques, sur la base d'isomorphismes de modèles. Sur la base de tels isomorphismes étendre les algorithmes élaborés pour les modèles structuraux aux analyses de séquences (notamment alignements), pour tenir compte de lois physiques réalistes.

-En parallèle des formulations algorithmiques, s'attaquer pour un certain nombre de modèles d'intérêt à la détermination des lois physiques réalistes (souvent non connues, du fait de l'adoption de lois simplifiées pour des motifs d'algorithmique).

-Avec des traitements algorithmiques efficaces, poursuivre et développer les analyses génomiques sur des bases structurales (notamment en termes d'identification de gènes), alors que la plupart des analyses actuelles reposent sur des bases 'linguistiques' et 'textuelles'.

-Développer les interfaces entres les analyses génomiques 'classiques', à grandes échelles, et les analyses génomiques basées sur des propriétés structurales..

PROJET

Algorithmique efficace pour des modèles réalistes en bioinformatique et biophysique. Détermination des lois physiques réalistes. Analyses

génomiques sur des bases structurales.

RESULTATS

-Détermination de la physique pour un modèle réaliste de transitions hélice-pelote dans les molécules d'ADN topologiquement contraintes [1]

-Algorithmes efficaces pour les alignements de séquences avec des lois physiques réalistes pour les gaps, par le biais d'isomorphismes avec les modèles structuraux [2]

-Modèles physiques d'ARN avec pseudo-nœuds [3]

-Extensions des analyses génomiques sur des bases physiques [4]

-Arbres et phylogénies sur la base d'analyses comparatives des génomes complets [5]

CALENDRIER

2005 Avril/Mai Démarrage officiel (notification de crédits)
2006 Réunion générale (Paris)

LABORATOIRES IMPLIQUES

Unité de Bio-Informatique Structurale, Institut Pasteur
Service de Physique Theorique, CE-Saclay, CEA
Unité de Génétique Moléculaire des Levures, Institut Pasteur
Groupe Logiciels et Banques de données, Institut Pasteur

MOTS CLES

Algorithmes efficaces et complexités algorithmiques. Modèles physiques réalistes. Alignements de séquences. Analyses génomiques et modèles structuraux. Phylogénies et comparaisons de génomes.

PUBLICATIONS PRINCIPALES

[1] T. Garel, H. Orland, E. Yeramian. Physical model for helix-coil transitions in supercoiled DNA. *Europhys. Lett.*, 2005 (soumis, en révision).

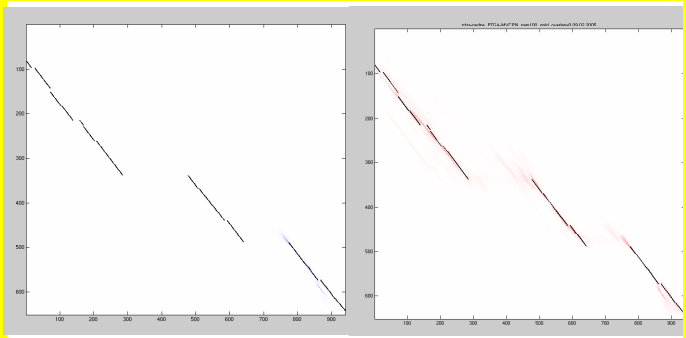
[2] E. Yeramian, E. Debonneuil. Physics of DNA and statistical mechanics of sequence alignments: realistic models with fast algorithms. Manuscrit, 2005 (sur le point d'être soumis).

[3] G. Vernizzi, H. Orland, A. Zee. Enumeration of RNA structures by matrix models. *Phys Rev Lett.*, 2005; 94(16):168103.

[4] E. Yeramian. Physics of DNA and genomics of *Apicomplexa*. Manuscrit, 2005 (sur le point d'être soumis).

[5] F. Tekaia, E. Yeramian. Genome Trees Based on Conservation Profiles: Phylogeny and Genome Dynamics. Plos Comput. Biol., 2005 (soumis, en révision).

Alignements de séquences et gaps réaliste



Sur la base d'isomorphismes entre modèles structuraux et modèles d'analyses de séquences, les idées algorithmiques élaborées au départ pour les modèles structuraux peuvent être étendues aux alignements de séquences. Dans un cadre probabiliste (mécanique statistique) il devient alors possible d'implémenter des alignements de séquences tenant compte de modèles réalistes pour les gaps (lois non-linéaires, et possibilités de gaps 'chevauchants'), avec des algorithmes efficaces en temps calcul. Pour des alignements de plus en plus difficiles les modèles réalistes (figure de droite, points en rouge) permettent d'obtenir des alignements (tests avec des cas connus structuralement, 'dot-plots' en noir), qui ne peuvent être obtenus avec les modèles simplifiés couramment utilisés aujourd'hui (figure de gauche, points en bleu, lois de pénalisations affines)