Diss. ETH No. 18177

Comparative Genomics Using Pairwise Evolutionary Distances

A dissertation submitted to the
ETH Zurich

for the degree of
Doctor of Sciences

presented by
Christophe Dessimoz
Dipl. Natw. ETH

born on 21 November 1980
citizen of Conthey (VS) and Vétroz (VS), Switzerland

accepted on the recommendation of
Prof. Dr. Gaston H. Gonnet, examiner
Prof. Dr. Amos Bairoch, co-examiner
Prof. Dr. Niko Beerenwinkel, co-examiner
Prof. Dr. Jörg Stelling, co-examiner
Prof. Dr. Martin Vingron, co-examiner

2009

Abstract

Comparative genomics is the study of genome structure, function, and evolution across species enabled by the recent availability of complete genome sequences. In this context, and assuming a common origin, one of the fundamental notion is the degree of evolutionary relatedness among sequences. This is usually expressed either as a tree structure, or as the matrix of distances between pairs of sequences. In this thesis, we explore and develop comparative genomics methods that rely on pairwise distances.

Part one of the thesis reviews and extends methods to estimate and combine distances between pairs of biological sequences. To estimate distances, the homologous characters must be identified; thus we begin with an evaluation of sequence alignment methods using simulation and real biological sequences. The main other contribution of part one consists in two methods to estimate the covariance of pairwise distances obtained from independent alignments (as can be the case in large-scale studies). The first method approximates the variance of the difference between two distances under empirical amino-acid substitution models. This is equivalent to computing the covariance between two distances that involve a common sequence. The second method is more general in the sense that it works with arbitrary Markovian models, and it also estimates the covariance of distances involving four distinct sequences, thereby allowing for the computation of full covariance matrices.

Part two applies these tools to large-scale comparative genomics analyses. The first application is an algorithm to detect differential gene losses, which often lead to misclassifications in orthologs detection methods based on genome-specific best hits, and in particular in the COGs database. Using it, we show that about a third of all COG groups include non-orthologs. These results motivate the second application, our own orthology inference effort, OMA. Besides the ability to detect differential gene losses, OMA has several other distinctive features: it computes evolutionary distances from pairwise maximum likelihood alignments, it uses confidence intervals to account for statistical inference uncertainty and to allow for many-to-many orthology, and clusters groups of orthologs using an edge-weighted clique algorithm. Our results are validated and compared with those resulting from 11 other projects and methods. The results of OMA are among the best in the phylogenetic tests. In functional tests, OMA, strict by design, performs comparatively well if high functional specificity is required. In terms of size, OMA is by a wide margin the largest effort of orthology inference; as of August 2008, we had performed all-against-all alignments of 670 complete genomes and had built orthologous matrices for the major clades of genomes. The third application of part two is an algorithm to detect lateral gene transfer based on pairwise distances. At its core, it computes a likelihood ratio between hypotheses of LGT and no LGT. As opposed to explicit phylogenetic LGT detection approaches, it avoids the high computational cost and

pitfalls associated with gene tree inference and reconciliation, while maintaining the high level of characterization (species involved in LGT, direction, distance to the LGT event in the past) associated with such methods. The results and validation of the algorithm, both through simulated and real data, show that the method outperforms common LGT detection approaches.

Résumé

Le génome est l'ensemble du matériel génétique d'un organisme encodé dans son ADN, et la génomique comparative est l'étude de la structure, de la fonction et de l'évolution des génomes au travers des espèces, rendue possible par le récent séquençage de génomes entiers. Dans ce contexte, et en supposant une origine commune à toutes les espèces, l'une des notions fondamentales c'est le degré de parenté entre espèces. Celle-ci s'exprime généralement soit sous forme d'arbre phylogénétique, soit sous forme de matrice de distances entre paires de séquences. Dans la présente thèse, nous explorons et développons particulièrement les méthodes de génomique comparative qui se basent sur les distances entre paires. Nous le faisons en deux parties principales.

Dans la première partie du mémoire, nous passons en revue et développons les méthodes actuelles pour l'estimation de distances entre paires de séquences biologiques. Afin d'estimer ces distances, les caractères homologues doivent au préalable être identifiés; nous commençons donc par une évaluation des méthodes d'alignement de séquence, d'abord en simulation puis par référence à des séquences biologiques réelles. Notre principale autre contribution dans cette première partie comprend deux méthodes pour estimer la covariance de distances obtenues en partant de paires alignées indépendamment les unes des autres (comme c'est souvent le cas dans les études génomiques à large échelle). La première méthode calcule la variance de la différence entre deux telles distances pour des modèles empiriques de substitution de caractères. Ceci équivaut à l'estimation de la covariance entre distances qui impliquent une séquence commune. La deuxième méthode est plus générale dans le sens qu'elle fonctionne pour tous les modèles Markoviens; elle permet aussi l'estimation de covariance entre deux distances qui impliquent quatre séquences distinctes, et peut ainsi s'employer pour calculer des matrices de covariance.

La seconde partie de la thèse consiste à appliquer ces méthodes dans le cadre d'études de génomique comparative à large échelle. La première application est un algorithme pour détecter des cas de perte différentielle de gènes, qui résultent souvent en des erreurs de classification au niveau de la prédiction d'orthologues, comme en témoigne le taux élevé d'erreurs dans la base de donnée COGs. En effet, nous démontrons la présence de séquences non-orthologues dans environ un tiers des groupes COG. Ce résultat motive la deuxième application, c'est-à-dire le développement de notre propre système d'identification d'orthologues, OMA. Mis à part sa capacité à détecter les pertes différentielles de gènes, OMA innove à plusieurs autres niveaux: l'algorithme se base sur les distances évolutives entre paires de séquences au lieu de scores souvent autrement utilisés par ailleurs; il emploie des intervalles de confiance pour tenir compte de l'incertitude statistique sur les distances estimées et pour permettre l'inférence de relations orthologues n-à-n ("many-to-many"); le regroupement d'orthologues se réalise par une approche de clique maximale pondérée. Nos résultats sont validés et comparés

avec ceux de 11 autres projets et méthodes. Les résultats d'OMA sont parmi les meilleurs dans les tests phylogénétiques. En terme de conservation de fonction, OMA, strict de par sa conception, assure de bonnes performances dans les cas où une spécificité fonctionnelle élevée est requise. En terme de taille, OMA recouvre de loin le plus grand nombre de génomes séquencés: au mois d'août 2008, nous avions aligné les gènes de 670 génomes complets, et construit des matrices d'orthologues pour les toutes les clades majeures. Enfin, la troisième application de cette deuxième partie est un algorithme pour détecter des transferts de gènes latéraux ("LGT") sur la base de distances des paires. En bref, il calcule un rapport de vraisemblance entre les hypothèses "transfert" et "pas de transfert". Contrairement aux approches phylogénétiques explicites, il évite les coûts de calcul élevés et les difficultés associées à l'inférence et la réconciliation d'arbres de gènes, tout en gardant leurs avantages en terme de caractérisation du cas en question (espèces impliquées dans le LGT, direction, distance évolutive au transfert). Les résultats et la validation de l'algorithme, au travers de données simulées et réelles, démontrent que cette méthode surpasse les méthodes standards de détection de LGT.