The WABI 2016 poster session is on Tuesday, Aug 23, starting at 15:30

Training alignment parameters for PacBio, Ion Torrent and Nanopore sequencers with last-train

Authors: Michiaki Hamada, Waseda University; Yukiteru Ono, IMSBIO; Kiyoshi Asai, University of Tokyo; Martin C. Frith, University of Tokyo

Contact: mhamada@waseda.jp

Keywords: alignment parameters, pair HMM, Nanopore, PacBio

Abstract:

Motivation: Recently, various new types of DNA sequencer have appeared, including Ion Torrent, PacBio RS, and Oxford Nanopore. These often have significant error rates, and each technology has its own characteristic error patterns (e.g. PacBio RS produces more insertions than deletions). It is often important to align such sequences to reference genomes as accurately as possible, but most recent aligners neglect the statistical model implicit in substitution and gap scores. It is likely that accuracy is maximized by using scores that fit the substitution and gap frequencies in the data. Results: Here, we firstly introduce a method to automatically train substitution and gap scores from a set of reads generated by an arbitrary sequencer. Our method is based on the Baum-Welch algorithm accelerated in combination with the X-drop algorithm, and handles insertion and deletion costs separately; The method is implemented in a general purpose aligner, LAST, as last-train. Secondly, we trained alignment parameters for PacBio RS, Ion Torrent and Nanopore sequencers, and performed several computational experiments with the trained parameters. Thirdly, to show a benefit of trained parameters, we performed haplotype phasing with Oxford Nanopore's long reads, indicating that trained parameters mitigate a pitfall of reference bias, reported in a previous study.

LAST: statistically rigorous, large-scale sequence comparison

Author: Martin C. Frith, AIST & University of Tokyo

Contact: mcfrith@edu.k.u-tokyo.ac.jp

Keywords: alignment, fastq, HMM, frameshift, rearrangement, trans-splicing

Abstract:

This poster presents LAST, open-source software for general-purpose, large-scale sequence comparison and alignment (http://last.cbrc.jp/).

- It is the only aligner that combines a traditional substitution score matrix (which models sequence divergence) with per-base uncertainty (e.g. fastq) in a rigorous way. This is useful for alignments with non-negligible divergence (e.g. cross-species alignment, ancient DNA) or unusual base frequencies (e.g. malaria, bisulfite converted DNA).
- It uses the statistical (pair HMM) basis of alignment to annotate the reliability of every column in an alignment.
- It is the only tool that can align DNA to proteins, *allowing frameshifts*, for genome-scale data. This is useful for: annotating pseudogenes, and analyzing metagenomic DNA (where frameshifts are surprisingly common).
- It can do "split alignment" of a query sequence to a genome, where it looks for a unique best match for each part of the query. It rigorously calculates the reliability (uniqueness) of each part of the alignment. This is useful for: cancer (DNA reads that cross rearrangement breakpoints), spliced RNA (where it models splice signals, intron sizes, and allows trans-splicing), and whole genome comparison (where different parts of one query chromosome match different parts of the target genome).
- Using newly-optimized transition seeds, LAST found ~20,000 new alignments between the human and mouse genomes, which are missing in the standard UCSC genome alignments.

Generalized haplotype consistency problem and characterization of haplotype consistent graphs

Authors: Bjarni Halldorsson, Reykjavik University; Juris Viksna, Institute of Mathematics and Computer Science, University of Latvia

Contact: juris.viksna@lumii.lv

Keywords: Haplotype consistency graphs, parameterized algorithms

Abstract:

Clark's graphs are used to describe consistency of haplotypes for a given set of individuals. The vertices of Clark's graph correspond to distinct individuals and two vertices are connected by an edge if these individuals share a distinctive haplotype tract. There are some variants of exact formalization of haplotype consistency, however the most often a graph G is called haplotype consistent if there is an assignment of two natural numbers to each of G vertices such that for each clique C in G there is exactly one number shared by all vertices of C.

Here we consider a natural generalization of this consistency requirement by assigning to each vertex of G two n-tuples of natural numbers. In this case a graph G is called haplotype weakly (strongly) consistent if there is such an assignment that for no two vertices the same pairs of n-tuples are assigned and that for each clique C in G and for each i=1…n there is a number that for each vertex v of C is equal to the i-th element of (exactly) one of the n-tuples assigned to v.

For n=1 both (weak and strong) consistency notions simply give us the most commonly used consistency requirement. However larger values of n reasonably well describe situations when we have only partial haplotype information – e.g. n=2 well represents the situation when we know that an individual on two different positions of a particular pair of chromosomes correspondingly has pairs of nucleotides {s,r} and {x,y}, but do not know whether r belongs to the same chromosome as x or to the same chromosome as y.

For case n=1 we have previously shown that G is not haplotype consistent if and only if G contains 3-claw or K4- (K4 with a single edge removed) as an induced subgraph. The number of non-isomorphic graphs grows very rapidly however with number of vertices, and we used a computer program to check for forbidden induced subgraphs for weak and strong consistency requirements for n=2.

For strong consistency the only forbidden subgraphs with no more than 7 vertices are K4- and 5-claw. This rises and interesting conjecture that G is not strongly haplotype n-consistent if and only if G contains (2^n+1)-claw or K4- as an induced subgraph. Unfortunately the proof technique for n=1 does not work for larger values of n and the conjecture remains an open problem.

For weak consistency there is 1 forbidden induced subgraph with 5 vertices, 9 with 6 vertices and 7 with 7 vertices. Currently we are analysing graphs with 8 vertices and are also working on optimization of computer program to be able to check for inconsistent graphs for the case n=3.

An alphabet-friendly privacy-preserving string search

Authors: Masanobu Jimbo, Waseda University; Hiroki Sudo, Waseda University; Koji Nuida, AIST; Kana Shimizu, Waseda University

Contact: jimwase@asagi.waseda.jp

Keywords: Biological Sequence Search, Privacy, Wavelet Matrix, BWT, Additively Homomorphic Encryption

Abstract:

Data sharing is one of the most promising approaches for accelerating life science, however, it also poses serious privacy risks. For example, the recent study pointed out even small amount of information leakage such as the 'beacon' of GA4GH's genome data-sharing project might cause a serious privacy breach. Similar problems are prevailing in many research domains, and it hinders access to many data resources potentially useful for a variety of scientific researches. To overcome the problem, it is necessary to develop a privacy preserving technique that enables an efficient secure database search. Among the previous methods for this purpose, PBWT-sec which is a combined algorithm of a cryptographic technique and positional Burrows Wheeler Transform is known to be efficient. The basic idea of the PBWT-sec is to compute a rank of the database string by using a cryptographic technique called recursive oblivious transfer. For a technical reason, computational complexity for calculating a rank is linear to alphabet size and a database string length. Therefore, the method works well only for a string with small alphabet size such as a DNA sequence, and it cannot handle an important life science data of large alphabet size such as a protein sequence and a time series data. To improve this drawback, we propose a novel method that achieves order of magnitude better computational performance by efficiently combining a wavelet matrix and the recursive oblivious transfer. The main problem of using the recursive oblivious transfer for computing rank is, that it requires a computational cost proportional to alphabet size S. To reduce the total cost, the proposed algorithm decomposes a query for computing a rank of a character c, into sub-queries each of which is for computing a rank of a bit of bit-decomposed c. Since a computational complexity for each sub-query does not depend on alphabet size and only depends on the size of the original string length N, the total computational complexity for calculating rank of c becomes O(N log S). The recursive oblivious transfer does not leak intermediate information, and thus the proposed method does not lose any advantage of the previous method in terms of security, while it largely increases the computational performance. We implemented the proposed method and confirmed that the observed CPU times for searching random strings with various alphabet sizes are concordant with the theoretical complexity. For example, the result of our method is around 100 times better than that of the previous method when the alphabet size is 4096. We also show results for searching on Pfam database and a real time series dataset. Those results support the efficiency of our approach, and we expect that the approach will be useful for wide range of life science.

Family-Free DCJ with Indels

Authors: Kevin Lamkiewicz, Center for Biotechnology, Genome Informatics, Bielefeld University; Pedro Feijao, Center for Biotechnology, Genome Informatics, Bielefeld University

Contact: klamkiew@cebitec.uni-bielefeld.de

Keywords: Rearrangements, Family-Free, Indel

Abstract:

Genome rearrangements are a widely studied field since they provide insight on how large scale structural evolution happens. Orthologous genes are as- signed to a gene family, indicating the similarity of these genes in terms of their function. In comparative genomics the assignment of gene families enables an easy comparison between genomes considering the orientation and order of their genes. Rearrangement events shuffle the genes of a genome in another order (and potentially orientation). The most common rearrangement model is the Double-Cut-and-Join (DCJ) model proposed by Yancopoulos et al. (2005). In order to calculate the DCJ-distance between two genomes, Bergeron et al. (2006) proposed the Adjacency Graph.

Recently Dörr et al. (2012) proposed the family-free model for comparative genomics, where the information of gene families is not included. This way the information of mutations, insertions and deletions at the sequence-level is considered as well. Martinez et al. (2015) adapted the family-free approach to the DCJ model. Due to the usage of the DCJ model, the family-free version is named the Family-Free Double-Cut-and-Join (FFDCJ) model. The FFDCJ model by Martinez et al. uses pairwise similarities between genomic sequences to build a bipartite graph, called the Gene Similarity Graph (GSG). The weighted adjacency graph (WAG) is derived from this graph by finding a maximum matching in the GSG. With the help of the WAG we are able to calculate the FFDCJ- distance. However, finding the maximum matching that minimizes this distance out of all possible matchings is shown to be NP-hard. The FFDCJ-distance problem has been solved with an integer linear program (ILP).

We propose an extension of the model and the ILP that considers insertion and deletion (indel) events. We introduce two different approaches that mea- sure the number of indel events – the Unitary model and the Block model. In the Unitary model each indel event is considered as a single operation. This model provides a proportional distance with regard to the evolutionary distance, however it overestimates the number of indel events. The Block model on the other hand gives a good estimation of indel events between two genomes, but underestimates the distance for two genomes that are far apart, i. e. that have a large evolutionary distance. We propose a post-modification of the FFDCJ- Indel distance, that combines the two indel models, to provide the advantages of both: the proportional distance and the satisfactorily estimation of indel events.

We present first results of our method by using simulated data from ALF (Dalquen et al., 2012). For this we simulated several datasets with increasing evolutionary distance, each with two genomes derived from Escherichia coli, that were affected by different rearrangements and sequence mutations. Using this simulations we observe that the Block model still overestimates the number of indel events. In our next step we want to adapt methods from the family- based approach to the FFDCJ-model, in order to give an even better result in terms of indel estimation.

Probabilistic Reconstruction of Ancestral Genomes using Intermediate Genomes

Authors: Kevin Lamkiewicz, Center for Biotechnology, Genome Informatics, Bielefeld University; Pedro Feijao, Center for Biotechnology, Genome Informatics, Bielefeld University

Contact: klamkiew@cebitec.uni-bielefeld.de

Keywords: Intermediate Genomes, Genome Reconstruction, DCJ, Maximum-Likelihood

Abstract:

The reconstruction of ancestral genomes is studied by evolutionary biologists for several years. It can be generalized to the Small Parsimony Problem (SPP), where a phylogenetic tree topology with extant genomes at its leaves is given. The aim is to reconstruct ancestor genomes that are located at the internal nodes of the given tree in a way, such that the overall distance of the tree is minimized.

Approaches to solve this can be divided into two groups – parsimonious approaches and homology approaches. The first one focuses on rearrangement events between two input genomes whereas the latter looks for conserved structures like common adjacencies or gene clusters. Examples for the event-based method are MGRA (Alekseyev, Pevzner, 2009), PATHGROUPS (Zheng, Sankoff, 2011) and GASTS (Xu, Moret, 2011). However, rearrangement distance problems are usually NP-hard for three or more input genomes. Therefore an exact implementation tends to be time consuming and not applicable for real data. Homology methods aim to use the information of conserved structures to assemble Contiguous Ancestral Regions (CARs) proposed by Ma et al. in 2006. This approach has been explored by many algorithms like PMAG (Yang et al., 2014), ProCARs (Perrin et al., 2014) and ANGES (Jones et al., 2012).

Feijao (2015) proposed an approach that is inspired by rearrangement models but not motivated as a distance-based method. Specifically he uses intermediate genomes (IGs) that arise from optimal rearrangement scenarios between two genomes. Common adjacencies between two genomes and ancestral adjacencies that are part of an IG can be found with the circular breakpoint graph. With his approach, Feijao extends the search space of homology methods by such ancestral adjacencies and restricts the search space of parsimonious methods.

We propose a probabilistic approach that is based on the Maximum-Likelihood (ML) method (Felsenstein, 1981). In our first step we only allow genomes with same gene content and size, i.e. no indel or duplication events. We sample IGs as proposed by Feijao (2015). Inspired by the work of Ma (2010) and Yang et al. (2014) we encode the presence of an adjacency as a binary variable. For each genome – the two input genomes and the sampled IG – a binary vector is created that stores the information of each binary variable. With this binary vector we infer the gene order of the potential ancestral genome. We then denote the IG with the highest probability as the common ancestor.

First results of our method will be presented. This includes the accuracy of correct gene order prediction and an evaluation with other tools. Next steps are the inclusion of extra information (i. e. an outgroup) and adapting the concept of duplication and indel events.

Protein Threading Optimization Using Consensus Homology Modeling

Authors: Tasmin Tamanna Haque, Bangladesh University of Engineering and Technology; Maliha Sarwat, Bangladesh University of Engineering and Technology； Swakkhar Shatabda, United International University； Mohammad Sohel Rahman, Bangladesh University of Engineering and Technology

Keywords: Protein structure prediction, Protein threading, Homology modeling

Contacts: sohel.kcl@gmail.com, swakkhar17@gmail.com, tangled27@gmail.com, sarwat.maliha@yahoo.com

Abstract:

This abstract represents ideas of threading optimization for predicting unknown protein structures. Protein Structure Prediction (PSP) problem is defined as the problem of determining the tertiary structure of a protein from knowledge of its primary structure and/or from knowledge of other structures (e.g., secondary structure components, templates from homologous proteins). We applied homology modelling in different sequences of unknown proteins and attempted to predict their structures based on the found homologs. The homologous proteins are obtained from servers like SPARKS- X/LOMETS and Protein Data Bank (PDB). Initially we applied sequential alignment (Needleman-Wunsch algorithm) between pairs of homologs and took the one closest to the target protein and labelled it as our consensus model. The templates that are matched with the target sequence generally helps to model positions of alpha helices and beta sheets. The other regions that might contain loops are modeled based on fragment assembly. Therefore the consensus model was further divided into several fragments of amino acids and optimized. Fragmentation causes each amino acid sequence to have several alternative arrangements and populate different structures. To evaluate these structures, we used scoring functions based on contact energies between residues, inspired by the robust geometric property that the distance between the carbon alpha atoms of two consecutive amino acids is preserved with a low variance. Another guiding principle of most prediction models is that protein native structure is thermodynamically stable and therefore it is at a free energy minimum. As a consequence, the problem of finding the native structure mainly relates to the representation of a protein by a model which allows the definition of a (free) energy for each of its conformations. Therefore we took help of Barrera et al contact energy matrix and evaluated different structures formed after fragmentation in consensus. Using the matrix, for each amino acid pair in contact, an (additive) energy is assigned and the energy of the protein can be estimated from the sum of all pairwise energies. Thus, a good structure can be predicted by minimum sum of energy. This idea summarizes in predicting unknown structured proteins with the help of their homologs and optimizing further by fragment assembly and preferable structure conformations guided by contact energy matrix. Our extended approach relates to rotating two different homologous proteins so that they overlap on mutually aligned portion and produce a relative structure to best match the target protein.

Adaptive local realignment via parameter advising

Authors: Dan DeBlasio and John Kececioglu, University of Arizona

Contact: deblasio@cs.arizona.edu

Keywords: Multiple sequence alignment, parameter advising, accuracy estimation

Abstract:

Mutation rates can vary across the residues of a protein, but when multiple sequence alignments are computed for protein sequences, typically the same choice of values for the substitution score and gap penalty parameters is used across the entire protein. We provide for the first time a new method called adaptive local realignment, which computes protein multiple sequence alignments that automatically use diverse alignment parameter settings in different regions of the input sequences. This allows the aligner's parameter settings to locally adapt across a protein to more closely follow varying mutation rates.

Our method builds on the Facet alignment accuracy estimator, and our prior work on global alignment parameter advising. In a computed alignment, for each region that has low estimated accuracy, a collection of candidate realignments is generated using a set of alternate parameter choices. If one of these alternate realignments has higher estimated accuracy than the original subalignment, it is replaced.

Adaptive local realignment significantly improves the quality of alignments over using the single best default parameter choice. In particular, local realignment, when combined with existing methods for global parameter advising, boosts alignment accuracy by almost 24% over the best default parameter setting on the hardest-to-align benchmarks.

A new version of the Opal multiple sequence aligner that incorporates adaptive local realignment, using Facet for parameter advising, is available free for non-commercial use at http://facet.cs.arizona.edu. This site also contains the benchmarks from our experiments, and optimal sets of parameter choices.

A preprint of this work is available on bioRxiv: D. DeBlasio and J. Kececioglu. Boosting alignment accuracy through adaptive local realignment. bioRxiv, doi:10.1101/063131, 2016.