Logiciels de Méthodologie et Statistiques

- Conception d’eCRF et data management : Clinsight
- Biostatistiques : nQuery (calcul d’échantillon nécessaire dans les études), SAS (analyses biostatistiques), Pest.
- Modélisation : logiciels Monolix (MOdèles NOn LInéaires à effets miXtes, MONOLIX group, Orsay, France) et R (R Development Core Team) pour les études de modélisation (pharmacocinétique et pharmacodynamique).
- KEM® : Logiciel propriétaire basé sur une exploration systématique et non-supervisée des règles logiques entre combinaisons de variables extraits d’une base de données. Ce logiciel se décline en plusieurs modules spécifiques métiers.
  - KEM® Biomarkers : Dans le domaine des biomarqueurs, identification de signatures prédictives à partir d’un grand nombre de variables pouvant être hétérogènes (génomiques, protéomiques, données cliniques …).
  - KEM® Clinical : Dans le cadre des études cliniques, identification des caractéristiques des patients répondeurs à un traitement ou des populations à risque afin d’optimiser le design des essais par la modification des critères d'inclusion/exclusion (stratification de population).
- R et SPSS : utilisé pour les statistiques descriptives (ACP, clustering hiérarchique) et inferentielles (tests statistiques),
- Data mining : SVM, Random Forest, Neural Network ; application à l'analyse de puces : normalisation, analyse différentielle.
- PLINK : pour les études génétiques (GWAS), étude cas-contrôle.
- Knime : outil de pipelining où nos propres modules sont développés pour s’assurer de la qualité et de la reproductibilité des analyses statistiques.
- Plateforme d’analyse de séquences partielles de métagénomes ; approches phylogénétiques et tests de diversité.
- PFGE pour profilage génétique.
- Plateforme classique d’analyses de séquences d’ADN.
- Profilage métabolique et identification de bactéries : approches omnilog, bioscreen & VITEK2, etc.
- Appareils PCR et PCRq pour identification de bactéries.
- Logiciels SAS, SPAD, développement spécifiques sous R.
- Suite de programmes pour la fouille de données symboliques : Coron
Logiciels de la plateforme informatique de l'INRA

A combined transmembrane topology and signal peptide prediction method.

A DNA Sequence Submission and Update Tool

A Java application/applet to display .scf traces and phred quality values.

A Sequence Viewer for basic bioinformatics. CLC Sequence Viewer creates a software environment enabling users to make a large number of bioinformatics analyses, combined with smooth data management, and excellent graphical viewing and output options.

A set of sequence comparison tools (fasta36, ggsearch...) used for alignment and database searching. For example, fasta compares a protein sequence to another protein sequence or to a protein database, or a DNA sequence to another DNA sequence or a DNA library.

A tandem repeat in DNA is two or more adjacent, approximate copies of a pattern of nucleotides. Tandem Repeats Finder is a program to locate and display tandem repeats in DNA sequences.

ABySS is a de novo, parallel, paired-end sequence assembler that is designed for short reads. The single-processor version is useful for assembling genomes up to 100 Mbases in size. The parallel version is implemented using MPI and is capable of assembling larger genomes.

Accurate Contact Prediction from large protein alignments

ACNUC allows to select sequences from many criteria from these three databases, to translate protein-coding genes in protein, and to extract selected sequences in user files. ACNUC is very efficient in providing direct access to coding regions (e.g. protein coding regions, tRNA or rRNA coding regions) of DNA fragments present in GenBank.

Agmial est une chaîne d’annotation de génomes microbiens, formée de deux modules indépendants. Le premier gère les séquences protéiques, le second les séquences nucléiques. Agmial soutient le principe que l’expert humain doit être placé au centre du processus d’annotation. Afin d’aider les annotateurs dans cette tache complexe et coûteuse en temps, le système est conçu pour automatiser au maximum le processus d’annotation et fournir des interfaces conviviales. Il implémenté une stratégie d’annotation. Le système est capable de travailler sur des séquences non finies (draft) et il permet l’annotation collaborative par des équipes d’annotateurs. Il est basé sur des standards informatiques (services web, système de gestion de base de données relationnelles, Java, ...) et bioinformatiques. Le système est distribué sous licence GPL. Agmial est actuellement utilisé par plusieurs laboratoires de l’INRA pour l’annotation ou la réannotation de génomes d’interêt agro-alimentaire.

align et align0 calcule un alignement global de deux sequences.

ALLPATHS-LG is a de Bruijn graph-based de novo assembler for large (and small) genomes.
ALLPATHS-LG is being developed by scientists at the Broad Institute.

An Improved Hidden Markov Model for Transmembrane Protein Topology Prediction and Its Applications to Complete Genomes
AnovArray permet la quantification des facteurs biologiques et des biais techniques, ainsi que l'identification des gènes différemment exprimés entre plusieurs conditions expérimentales (deux et plus) pour des expériences transcriptomiques issues de macroarray et microarray dans la cadre d'un plan d'expérience factoriel équilibré et d'un modèle complet. Ce package est développé en SAS (logiciel statistique) et bénéficie en conséquence de toutes les procédures statistiques de ce logiciel. Les méthodes statistiques dans ce package sont l'analyse de la variance (ANOVA) et les tests multiples de type FDR (False Discovery Rate).

Apollo is a genomic annotation viewer and editor. There are currently two branches of Apollo, one primarily used for genome browsing and maintained at Ensembl, and the other primarily used for genome annotation and maintained at the Berkeley Drosophila Genome Center. The latter is part of the GMOD project.

Arachne is a tool for assembling genome sequences from whole genome shotgun reads, mostly in forward-reverse pairs obtained by sequencing clone ends.

Artemis is a free genome viewer and annotation tool that allows visualization of sequence features and the results of analyses within the context of the sequence, and its six-frame translation.

AutoDock Vina is a new program for drug discovery, molecular docking and virtual screening, offering multi-core capability, high performance and enhanced accuracy and ease of use.

BFAST : Blat-like Fast Accurate Search Tool BFAST facilitates the fast and accurate mapping of short reads to reference sequences. Some advantages of BFAST include: * Speed: enables billions of short reads to be mapped quickly. * Accuracy: A priori probabilities for mapping reads with defined set of variants. * An easy way to measurably tune accuracy at the expense of speed.

BioArray Software Environment (BASE) est une base de données permettant de gérer l'importante quantité de données générées par des analyses de bio-puces. BASE gère les informations biologiques, les données brutes et les images. BASE possède également des outils de normalisation, de visualisation et d'analyse des données.

BLAT is a DNA/Protein Sequence Analysis program written by Jim Kent at UCSC. It is designed to quickly find sequences of 95% and greater similarity of length 40 bases or more. It may miss more divergent or shorter sequence alignments. It will find perfect sequence matches of 33 bases, and sometimes find them down to 22 bases. BLAT on proteins finds sequences of 80% and greater similarity of length 20 amino acids or more. In practice DNA BLAT works well on primates, and protein blat on land vertebrates.

Bowtie is an ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of over 25 million 35-bp reads per hour. Bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small: typically about 2.2 GB for the human genome (2.9 GB for paired-end).

BWA is a fast light-weighted tool that aligns short sequences to a sequence database, such as the human reference genome. By default, BWA finds an alignment within edit distance 2 to the query sequence, except for disallowing gaps close to the end of the query. It can also be tuned to find a fraction of longer gaps at the cost of speed and of more false alignments.
CCP4 exists to produce and support a world-leading, integrated suite of programs that allows researchers to determine macromolecular structures by X-ray crystallography, and other biophysical techniques. CCP4 aims to develop and support the development of cutting edge approaches to experimental determination and analysis of protein structure, and integrate these approaches into the suite. CCP4 is a community based resource that supports the widest possible researcher community, embracing academic, not for profit, and for profit research. CCP4 aims to play a key role in the education and training of scientists in experimental structural biology. It encourages the wide dissemination of new ideas, techniques and practice.

CD-HIT stands for Cluster Database at High Identity with Tolerance. The program (cd-hit) takes a fasta format sequence database as input and produces a set of 'non-redundant' (nr) representative sequences as output.

cd-hit-454

Celera Assembler is scientific software for DNA research. It can reconstruct long sequences of genomic DNA from the fragmentary data produced by whole-genome shotgun sequencing. The Celera Assembler is mature, efficient, open-source software written mostly in C for unix operating systems.

CENSOR is a software tool which screens query sequences against a reference collection of repeats and "censors" (masks) homologous portions with masking symbols, as well as generating a report classifying all found repeats.

CGView is a Java package for generating high quality, zoomable maps of circular genomes. Its primary purpose is to serve as a component of sequence annotation pipelines, as a means of generating visual output suitable for the web. Feature information and rendering options are supplied to the program using an XML file, a tab delimited file, or an NCBI ptt file. CGView converts the input into a graphical map (PNG, JPG, or Scalable Vector Graphics format), complete with labels, a title, legends, and footnotes. In addition to the default full view map, the program can generate a series of hyperlinked maps showing expanded views. The linked maps can be explored using any web browser, allowing rapid genome browsing, and facilitating data sharing. The feature labels in maps can be hyperlinked to external resources, allowing CGView maps to be integrated with existing web site content or databases. For examples of the various output types, see the CGView gallery.

Circos is a software package for visualizing data and information. It visualizes data in a circular layout — this makes Circos ideal for exploring relationships between objects or positions. There are other reasons why a circular layout is advantageous, not the least being the fact that it is attractive.

Class2G permet de classer les gènes en deux groupes en utilisant un modèle de mélange. Les principales caractéristiques sont d'une part l'affectation des gènes est associée à une probabilité, et d'autre part l'analyse d'un macroarray est indépendante d'une référence. Class2G est intégrée au système BASE (BioArray Software Environment) par l'intermédiaire d'un plug-in perl, et est développé dans l'environnement statistique R. BASE permet d'accéder à une interface web conviviale, d'utiliser un seul environnement pour le stockage et l'analyse de données. Class2G a été utilisé pour la détection de gènes présents et absents de E. faecalis dans le cadre de l'analyse d'une trentaine de macroarray (P.Serror - INRA Jouy-en-Josas - UBLO).

Clustal Omega is the latest addition to the Clustal family. It offers a significant increase in scalability over previous versions, allowing hundreds of thousands of sequences to be aligned in only a few
hours. It will also make use of multiple processors, where present. In addition, the quality of alignments is superior to previous versions, as measured by a range of popular benchmarks.

CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing.

Consed/Autofinish is a tool for viewing, editing, and finishing sequence assemblies created with phrap. Finishing capabilities include allowing the user to pick primers and templates, suggesting additional sequencing reactions to perform, and facilitating checking the accuracy of the assembly using digest and forward/reverse pair information.

CONSEL is a program package consists of small programs written in C language. It calculates the probability value (i.e., p-value) to assess the confidence in the selection problem. Although CONSEL is applicable to any selection problem, it is mainly designed for the phylogenetic tree selection. CONSEL does not estimate the phylogenetic tree by itself, but CONSEL does read the output of the other phylogenetic packages, such as Molphy, PAML, PAUP*, TREE-PUZZLE, and PhyML. CONSEL calculates the p-value using several testing procedures; the bootstrap probability, the Kishino-Hasegawa test, the Shimodaira-Hasegawa test, and the weighted Shimodaira-Hasegawa test. In addition to these conventional tests, CONSEL calculates the p-value based on the approximately unbiased test using the multi-scale bootstrap technique. This newly developed method gives less biased results than the conventional methods.

Coot is for macromolecular model building, model completion and validation, particularly suitable for protein modelling using X-ray data. Coot displays maps and models and allows model manipulations such as idealization, real space refinement, manual rotation/translation, rigid-body fitting, ligand search, solvation, mutations, rotamers, Ramachandran plots, skeletonization, non-crystallographic symmetry and more.

Cross_Match uses the same algorithm as Swat but also allows the comparison of a pair of sequences to be constrained to bands of the Smith-Waterman matrix that surround one or more matching words in the sequences. This substantially increases speed for large-scale nucleotide sequence comparisons without compromising sensitivity.

Cufflinks assembles transcripts and estimates their abundances in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one.

cutadapt is used to remove adapter sequences from high-throughput sequencing data. This is usually necessary when the read length of the sequencing machine is longer than the molecule that is sequenced, for example when sequencing microRNAs.

Cytoscape is an open source bioinformatics software platform for visualizing molecular interaction networks and biological pathways and integrating these networks with annotations, gene expression profiles and other state data. Although Cytoscape was originally designed for biological research, now it is a general platform for complex network analysis and visualization. Cytoscape core distribution provides a basic set of features for data integration and visualization.

DOTUR est un programme qui prend en entrée une matrice décrivant les distances génétiques entre des séquences d'ADN pour les assigner à des unités taxonomiques opérationelles (OTUs). DOTUR utilise la composition des OTUs pour calculer des courbes de raréfaction et de collection.
ELPH is a general-purpose Gibbs sampler for finding motifs in a set of DNA or protein sequences. The program takes as input a set containing anywhere from a few dozen to thousands of sequences, and searches through them for the most common motif, assuming that each sequence contains one copy of the motif. We have used ELPH to find patterns such as ribosome binding sites (RBSs) and exon splicing enhancers (ESEs).

ESPript, Easy Sequencing in Postscript, is a utility to generate a pretty PostScript output from aligned sequences.

ESPRIIT is a pipeline for estimating species richness using large collections of 16S rRNA pyrosequences.

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

FigTree is designed as a graphical viewer of phylogenetic trees and as a program for producing publication-ready figures. As with most of my programs, it was written for my own needs so may not be as polished and feature-complete as a commercial program. In particular it is designed to display summarized and annotated trees produced by BEAST.

FinchTV (Finch Trace Viewer), a cross-platform graphical viewer for chromatogram files.

Findtarget est un outil de comparaison génomique qui permet de cibler des gènes d'intérêts chez un micro-organisme dont le génome est séquencé. Il utilise des données issues de blast.

FPC (fingerprinted contigs) is an interactive program for building contigs from fingerprinted clones, where the fingerprint for a clone is a set of restriction fragments.

FSA-BLAST is a new version of the popular BLAST (Basic Local Alignment Search Tool) bioinformatics tool, used to search genomic databases containing either protein or nucleotide sequences. FSA stands for Faster Search Algorithm; FSA-BLAST is twice as fast as NCBI-BLAST with no loss in accuracy.

GALF-P is a novel framework for TFBS identification (motif discovery) in DNA sequences. It consists of Genetic Algorithm with Local Filtering (GALF) and the post-processing procedure based on adaptive adding and removing. GALF-P achieves both effectiveness and efficiency, and provides reliable performance over the other state-of-art GA based approaches. The post-processing procedure is designed for zero or more TFBSs in each sequence.

GASSST : Global Alignment Short Sequence Search Tool * GASSST finds global alignments of short DNA sequences against large DNA banks. * GASSST strong point is its ability to perform fast gapped alignments. * It works well for both short and longer reads. It currently has been tested for reads up to 500bp. * The software is freely available for download under the CECILL version 2 License.
Gblocks

Genewise permet de comparer une protéine à une banque d’ADN et en prédire sa structure, tout en se déchargeant des problèmes liés au sequencing et d’introns.

GiMSAN (GibbsMarkov with Significance ANalysis): a novel tool for de novo motif finding. GiMSAN combines GibbsMarkov, our variant of the Gibbs Sampler, described here for the first time, with our recently introduced significance analysis.

Glimmer (Gene Locator and Interpolated Markov ModelER) prédit la position des gènes dans une séquence d’ADN (bactérie, arche, virus) en s’appuyant sur des modèles de Markov.

GMAP (genomic mapping and alignment program for mRNA and EST sequences): gmap, a standalone program for mapping and aligning cDNA sequences to a genome. The program maps and aligns a single sequence with minimal startup time and memory requirements, and provides fast batch processing of large sequence sets. The program generates accurate gene structures, even in the presence of substantial polymorphisms and sequence errors, without using probabilistic splice site models. GSNAP (Genomic Short-read Nucleotide Alignment Program): GSNAP implements computational methods for fast detection of complex variants and splicing in short reads, based on a successively constrained search process of merging and filtering position lists from a genomic index. It can align both single- and paired-end reads as short as 14 nt and of arbitrarily long length. It can detect short- and long-distance splicing, including interchromosomal splicing, in individual reads, using probabilistic models or a database of known splice sites.

G-Mo.R-Se is a method aimed at using RNA-Seq short reads to build de novo gene models. First, candidate exons are built directly from the positions of the reads mapped on the genome (without any ab initio assembly of the reads), and all the possible splice junctions between those exons are tested against unmapped reads : the testing of junctions is directed by the information available in the RNA-Seq dataset rather than a priori knowledge about the genome. Exons can thus be chained into stranded gene models.

Goby is a next-gen data management framework designed to facilitate the implementation of efficient next-gen data analysis pipelines. Goby provides compressed file formats that are time and space efficient. It also provides a few utilities that support the most common secondary data analyses

GRAIL is a suite of tools designed to provide analysis and putative annotation of DNA sequences both interactively and through the use of automated computation.

GRAPe is a tool for computing genome re-alignment using marginalized posterior decoding. To answer this question, GRAPe uses the Marginalized Posterior Decoding (MPD) algorithm which uses the posterior distribution of alignments to optimize the correct assignment of homology of individual nucleotides, instead of finding a single most probable alignment. Simulations show that the MPD algorithm has higher sensitivity and specificity than the Viterbi and Needleman-Wunsch algorithms.

grepseq

GRIL is a tool to detect the locations of genomic rearrangements in a set of sequences.
HHsearch is a software suite for detecting remote homologues of proteins and generating high quality alignments for homology modeling and function prediction.

HMMER: profile HMMs for protein sequence analysis Profile hidden Markov models (profile HMMs) can be used to do sensitive database searching using statistical descriptions of a sequence family's consensus.

Html4blast est un logiciel d'analyse et de présentation des résultats de Blast.

iCORN (iterative correction of reference nucleotides) can correct genome sequences with short reads. Reads are mapped iteratively against the genome sequences, so far by SSAHA. Discrepancies between the multiple alignments of the mapping reads and reference are corrected, if by the correction the amount of perfect mapping reads doesn't decrease.

Illumina CASAVA-1.8 FASTQ Filter

indel-Seq-Gen (iSG) is a biological sequence simulation program that simulates highly divergent DNA sequences and protein superfamilies. This is accomplished through the addition of subsequence length constraints and lineage- and site-specific evolution. iSG tracks insertion and deletion processes that occur during the simulation run. iSG records all evolutionary events and outputs the "true" multiple alignment of the sequences, and can generate a larger simulated sequence space by allowing the use of multiple related root sequences. iSG can be used to test the accuracy of multiple alignment methods, evolutionary hypotheses, ancestral protein reconstruction methods, and protein superfamily classification methods. iSG utilizes a highly modified version of the substitution engine from Seq-Gen v1.3.2.

InParanoid is a program for automatic identification of orthologs while differentiating between inparalogs and outparalogs. An InParanoid cluster is seeded by a reciprocally bestmatching ortholog pair, around which inparalogs are gathered independently, while outparalogs are excluded. The InParanoid database is a collection of pairwise ortholog groups aiming to include all 'completely sequenced' eukaryotic genomes. By this we mean above 6X coverage, and less than 1% X letters in the protein sequences.

Jalview is a multiple alignment editor

JELLYFISH is a tool for fast, memory-efficient counting of k-mers in DNA. A k-mer is a substring of length k, and counting the occurrences of all such substrings is a central step in many analyses of DNA sequence. JELLYFISH can count k-mers using an order of magnitude less memory and an order of magnitude faster than other k-mer counting packages by using an efficient encoding of a hash table and by exploiting the "compare-and-swap" CPU instruction to increase parallelism. JELLYFISH is a command-line program that reads FASTA and multi-FASTA files containing DNA sequences. It outputs its k-mer counts in an binary format, which can be translated into a human-readable text format using the "jellyfish dump" command. See the documentation below for more details.

La démarche centrale est basée sur les techniques émanant de l'apprentissage automatique (classification) et le traitement automatique des langues mais aussi d'une méthode sociologique appelée GST (Graphe SocioTechnique) de façon a construire des indices d'évolution de l'innovation grâce à la terminologie utilisée au cours du temps.

LALNVIEW is a graphical program for visualizing local alignments between two sequences (protein or nucleic acids) [reference]. Sequences are represented by colored rectangles to give an overall
picture of the similarities between the two sequences. Blocks of similarity between the two sequences are colored according to the degree of identity between the two segments.

Le paquetage pftools est une collection de programmes expérimentaux qui permet de manipuler le format généralisé de profils et implémente les méthodes de recherche de PROSITE. Les commandes accessibles sont les suivantes : gtop, pfsearch, pfsan, psa2msa, pfmake, pfw, ptoh, htop, pftscale, pftof.

MAFFT is a multiple sequence alignment program for unix-like operating systems. It offers a range of multiple alignment methods, L-INS-i (accurate; for alignment of <7200 sequences), FFT-NS-2 (fast; for alignment of <70,000 sequences), etc.

Mapsembler is a targeted assembly software. It takes as input a set of NGS raw reads and a set of input sequences (starters). It first determines if each starter is read-coherent, e.g. whether reads confirm the presence of each starter in the original sequence. Then for each read-coherent starter, Mapsembler outputs its sequence neighborhood as a linear sequence or as a graph, depending on the user choice.

MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery MapSplice est un algorithme de seconde génération de détection de sites d'épissage alternatifs. Son objectif est de détecter les sites d'épissage de façon sensible et spécifique en maintenant une bonne efficacité au niveau CPU et mémoire. MapSplice peut être appliqué aux reads courts (>75 pb) et long (75 pb). Il ne dépend ni des caractéristiques du site d'épissage ni de la longueur de l'intron, par conséquent, il peut détecter de nouveaux sites canoniques et non-canoniques d'épissage. MapSplice s'appuie sur la qualité et la diversité d'alignements des reads pour augmenter la précision de détection des sites d'épissage.

Maq is a software that builds mapping assemblies from short reads generated by the next-generation sequencing machines. It is particularly designed for Illumina-Solexa 1G Genetic Analyzer, and has a preliminary functionality to handle AB SOLiD data.

Mascot est un outil de recherche puissant qui utilise des données de spéctrométrie de masse pour identifier des protéines à partir de séquences primaires des bases de données.

Matrix2png is a simple but powerful program for making visualizations of microarray data and many other data types. It generates PNG formatted images from text files of data. It is fast, easy to use, and reasonably flexible. It can be used to generate publication-quality images, or to act as a image generator for web applications. Our group has found it useful for imaging all kinds of matrix-based data, not just microarray data.

mauve

MEGAN - Metagenome Analysis Software

MERLIN est un package qui permet d’effectuer des analyses génétiques rapides de pedigrees (analyses de liaison, d’association, haplotypes...).

MetaGeneAnnotator

MetaSim - A Sequencing Simulator for Genomics and Metagenomics
MGLTools is a software developed at the Molecular Graphics Laboratory (MGL) of The Scripps Research Institute for visualization and analysis of molecular structures. Short description and demo of its three main applications are given below. Navigation portlet on the left has links to downloads, screenshots, documentation section of this website where you can find more information about MGLTools. Please visit MGL Bugzilla to submit a bug report or to request a new feature.

MIRA is a Whole Genome Shotgun and EST Sequence Assembler for Sanger, 454 and Solexa / Illumina. It can perform Hybrid de-novo assemblies as well as SNP and mutations discovery for mapping assemblies.

MMSEQ: haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads pipeline The flowchart to the right depicts the MMSEQ pipeline for obtaining expression estimates from RNA-seq data. There are two routes, with starting points labelled A and B. Route A is quite fast and straightforward to run and uses pre-existing transcript sequences for alignment. Route B requires more time, as it involves the creation of custom transcript sequences based on the data.

MOCAT is a package for analyzing metagenomics datasets. Currently MOCAT supports Illumina single- and paired-end reads in raw FastQ format.

ModelGenerator is a model selection program that selects optimal amino acid and nucleotide substitution models from Fasta or Phylip alignments. ModelGenerator supports 56 nucleotide and 96 amino acid substitution models.

MODELLER is used for homology or comparative modeling of protein three-dimensional structures (1,2). The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms.

MolScript is a program for displaying molecular 3D structures, such as proteins, in both schematic and detailed representations.

MOSAIK is a reference-guided assembler comprising of four main modular programs: *
MosaicBuild * MosaicAligner * MosaicSort * MosaicAssembler. MosaicBuild converts various sequence formats into Mosaic’s native read format. MosaicAligner pairwise aligns each read to a specified series of reference sequences. MosaicSort resolves paired-end reads and sorts the alignments by the reference sequence coordinates. Finally, MosaicAssembler parses the sorted alignment archive and produces a multiple sequence alignment which is then saved into an assembly file format.

MPscan: fast localisation of multiple reads in genomes

Multiple Alignment with N Gapped OligossMANGO: A NEW APPROACH TO MULTIPLE SEQUENCE ALIGNMENT

Multiple sequence alignment program. It provides an integrated environment for performing multiple sequence and profile alignments and analysing the results.

Multipoint analysis of pedigree data including: non-parametric linkage analysis, LOD-score computation, information-content mapping, haplotype reconstruction
MUMmer is a system for rapidly aligning entire genomes, whether in complete or draft form. For example, MUMmer 3.0 can find all 20-basepair or longer exact matches between a pair of 5-megabase genomes in 13.7 seconds, using 78 MB of memory, on a 2.4 GHz Linux desktop computer. MUMmer can also align incomplete genomes; it can easily handle the 100s or 1000s of contigs from a shotgun sequencing project, and will align them to another set of contigs or a genome using the NUCmer program included with the system. If the species are too divergent for a DNA sequence alignment to detect similarity, then the PROmer program can generate alignments based upon the six-frame translations of both input sequences.

MUSCLE stands for MUltiple Sequence Comparison by Log-Expectation.

MView is a tool for converting the results of a sequence database search (BLAST, FASTA, etc.) into the form of a coloured multiple alignment of hits stacked against the query. Alternatively, an existing multiple alignment (MSF, PIR, CLUSTALW, etc.) can be processed. In either case, the output is simply HTML, so the result is platform independent and does not require a separate application or applet to be loaded. MView is NOT a multiple alignment program, nor is it a general purpose alignment editor.

Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy.

Nesoni focusses on analysing the alignment of reads to a reference genome. We use the SHRiMP read aligner, as it is able to detect small insertions and deletions in addition to SNPs. Nesoni can call a consensus of read alignments, taking care to indicate ambiguity. This can then be used in various ways: to determine the protein level changes resulting from SNPs and indels, to find differences between multiple strains, or to produce n-way comparison data suitable for phylogenetic analysis in SplitsTree4. Alternatively, the raw counts of bases at each position in the reference seen in two different sequenced strains can compared using Fisher's Exact Test.

NetLogo is a programmable modeling environment for simulating natural and social phenomena. It was authored by Uri Wilensky in 1999 and has been in continuous development ever since at the Center for Connected Learning and Computer-Based Modeling.

Newbler is a package of three data analysis applications made by Roche 454: the GS De Novo Assembler (with or without contig scaffolding using Paired End reads), the GS Reference Mapper, and the GS Amplicon Variant Analyzer (AVA). An additional application, the GS Run Browser, is an interactive Run browser/ troubleshooting tool which displays graphically the images, some intermediate data, and various output metrics from a sequencing Run. The software package also includes the SFF Tools commands for handling and using the data files (called Standard Flowgram Format or SFF files) that hold the sequencing trace data.

NJplot is a tree drawing program able to draw any binary tree expressed in the standard phylogenetic tree format (e.g., the format used by the PHYLIP package). NJplot is especially convenient for rooting the unrooted trees obtained from parsimony, distance or maximum likelihood tree-building methods.

Novoalign is an aligner for single-ended and paired-end reads from the Illumina Genome Analyser. Novoalign finds global optimum alignments using full Needleman-Wunsch algorithm with affine gap penalties.
novoSNP is a program that will help you find variations (SNPs and short INDELs) in resequencing projects. It takes a reference sequence and a number of sequencing trace files as input, and generates a list of possible variations with a quality score. novoSNP allows you to easily filter, sort and check the variations found visually and keep track of your verifications.

NUPACK is a growing software suite for the analysis and design of nucleic acid systems. The package currently enables thermodynamic analysis of dilute solutions of interacting nucleic acid strands, and sequence design for complexes of nucleic acid strands intended to adopt a target secondary structure at equilibrium. NUPACK algorithms are formulated in terms of nucleic acid secondary structure. In most cases, pseudo-knots are excluded from the structural ensemble. Much of this software may be conveniently run through the NUPACK web server at http://www.nupack.org (Zadeh et al., 2010b).

Obo-Edit est un éditeur d'ontologie dans le format obo. Le format obo a été défini originellement pour Gene Ontology et se répand dans la communauté bioinformatique. Quelques dizaines d'ontologies en format obo sont disponibles et éditable par Obo-Edit.

OCOUNT is a fast C command-line utility that has been written in the course of TETRA's development. It counts oligonucleotides in DNA sequences and computes Markov-Model-based z-scores.

OrthoMCl est un logiciel qui construit les clusters d'orthologue à partir de fichiers multifasta contenant des CDS.

Otterlace is an interactive, graphical client, which uses a local acedb database with Zmap and perl/Tk tools to curate genomic annotation. Annotation is stored in an extended Ensembl schema (the "otter" database), which presents the annotator with contiguous regions of a chromosome. The acedb database provides local persistent storage, so that if the software or desktop machine crashes, reboots or is exited, the editing session can be recovered. Since all communication goes through the Sanger web server, annotators can work wherever there is a network connection.

PatScan is a pattern matcher which searches protein or nucleotide (DNA, RNA, tRNA etc.) sequence archives for instances of a pattern which you input.

pfam_scan.pl - search protein fasta sequences against the Pfam library of HMMs.

Phd2fasta reads phd files and writes sequence and quality value FASTA files, which phrap and cross_match need as input. Phred and consed write sequence and quality value information in 'phd' output files. A phd file contains information in a header, the called bases, the base quality values, and the base call trace locations.

PHENIX is a software suite for the automated determination of macromolecular structures using X-ray crystallography and other methods.

phrap is a program for assembling shotgun DNA sequence data. Among other features, it allows use of the entire read and not just the trimmed high quality part, it uses a combination of user-supplied and internally computed data quality information to improve assembly accuracy in the presence of repeats, it constructs the contig sequence as a mosaic of the highest quality read segments rather than a consensus, it provides extensive assembly information to assist in trouble-
shooting assembly problems, and it handles large datasets.

Phred reads DNA sequencer trace data, calls bases, assigns quality values to the bases, and writes the base calls and quality values to output files. Phred can read trace data from chromatogram files in the SCF, ABI, and ESD formats. It automatically determines the file format, and whether the chromatogram file was compressed using gzip, bzip2, or UNIX compress. After calling bases, phred writes the sequences to files in either FASTA format, the format suitable for XBAP, PHD format, or the SCF format. Quality values for the bases are written to FASTA format files or PHD files, which can be used by the phrap sequence assembly program in order to increase the accuracy of the assembled sequence. phred, phrap, consed are Unix programs that work as a group for analysis of new DNA sequences. They do the following: phred: Base calling and quality assignments phrap: Contig formation and new quality assignments consed: Visual X-Windows graphic interface, to view and edit alignments and contigs, and to view the original traces

PHYLIP is a free package of programs for inferring phylogenies.

Phylogenetic Analysis by Maximum Likelihood - Logiciel de référence pour détecter une pression de détection positive sur certains sites.

PLAST : Parallel Local Alignment Search Tool

PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner.

PolyPhred is a program that compares fluorescence-based sequences across traces obtained from different individuals to identify heterozygous sites for single nucleotide substitutions.

Prediction of transmembrane helices and topology for transmembrane proteins using hidden Markov models

PRINSEQ CAN BE USED TO FILTER, REFORMAT, OR TRIM YOUR GENOMIC AND METAGENOMIC SEQUENCE DATA. IT GENERATES SUMMARY STATISTICS OF YOUR $ GRAPHICAL AND TABULAR FORMAT.

PROBCONS is a novel tool for generating multiple alignments of protein sequences. Using a combination of probabilistic modeling and consistency-based alignment techniques, PROBCONS has achieved the highest accuracies of all alignment methods to date. On the BALiBASE benchmark alignment database, alignments produced by PROBCONS show statistically significant improvement over current programs, containing an average of 7% more correctly aligned columns than those of T-Coffee, 11% more correctly aligned columns than those of CLUSTAL W, and 14% more correctly aligned columns than those of DIALIGN.

ProbeMatch is a sequence alignment program that finds sequence alignments for short DNA sequences (36-50 bp). Unlike other programs such as eland and soap that perform ungapped alignment allowing up to 2 substitution, Probematch performs *gapped* alignment, allowing up to 3 errors including substitution, insertion, and deletion.

procheck

Program MOLE is an universal toolkit for rapid and fully automated location and characterization of
channels, tunnels and pores in molecular structures. The core of MOLE algorithm is a Dijsktra path search algorithm, which is applied to a Voronoi mesh. MOLE is a powerful software (overcoming some limitations of CAVER tool) for exploring large molecular channels, complex networks of channels and molecular dynamics trajectories (AMBER ascii traj and parm7 are supported) in which analysis of a large number of snapshots is required.

Programme de comparaison et d'alignement des structures 3D des protéines. VAST est basé sur une procédure en 2 étapes. Dans la première étape on utilise une description simplifiée des protéines où les éléments de structure secondaire sont représentés par des vecteurs. Le but de cette première étape est de trouver le sous-ensemble des vecteurs qui se superimposent au mieux entre les 2 structures. La significativité du résultat est évaluée en calculant la probabilité d'observer cette superimposition juste par chance. Dans la seconde étape on revient à une description atomique des structures 3D en décrivant la chaîne polypeptique par les positions des CA de chaque résidu. L'objectif de cette seconde étape est d'établir une correspondance univoque (alignement) entre les CA jouant le même rôle dans les 2 structures. On cherche à obtenir l'alignement contenant les plus de paires de CA et le rmsd (root mean square deviation) le plus faible. Pour ce faire l'algorithme est amené à répondre à des questions comme : quel alignement, l'un comprenant 100 paires de CA et ayant un rmsd 3 A, et l'autre comprenant 60 paires de CA et un rmsd de 2 A est le meilleur? Ce problème est résolu en considérant l'alignement qui a la probabilité la plus faible d'être généré par hasard.

PSIPRED is a simple and reliable secondary structure prediction method, incorporating two feed-forward neural networks which perform an analysis on output obtained from PSI-BLAST (Position Specific Iterated - BLAST). Version 2.0 of PSIPRED includes a new algorithm which averages the output from up to 4 separate neural networks in the prediction process to further increase prediction accuracy.

PSORT is a computer program for the prediction of protein localization sites in cells. It receives the information of an amino acid sequence and its source origin, e.g., Gram-negative bacteria, as inputs. Then, it analyzes the input sequence by applying the stored rules for various sequence features of known protein sorting signals. Finally, it reports the possibility for the input protein to be localized at each candidate site with additional information.

Pymol est un logiciel de visualisation moléculaire associé à un interpréteur Python qui permet la visualisation en temps réel ainsi que la génération rapide et de qualité d’animations et d’images d’assemblages moléculaires.

PyNAST: Python Nearest Alignment Space Termination tool PyNAST is a reimplementation of the NAST sequence aligner, which has become a popular tool for adding new 16s rDNA sequences to existing 16s rDNA alignments. This reimplementation is more flexible, faster, and easier to install and maintain than the original NAST implementation. PyNAST is built using the PyCogent Bioinformatics Toolkit. The first versions of PyNAST (through PyNAST 1.0) were written to exactly match the results of the original NAST algorithm. Beginning with the post PyNAST 1.0 development code, PyNAST no longer exactly matches the NAST output but is instead focused on getting better alignments. Users who wish to exactly match the results of NAST should download PyNAST 1.0.

Quake is a package to correct substitution sequencing errors in experiments with deep coverage (e.g. >15X), specifically intended for illumina sequencing reads. Quake adopts the k-mer error correction framework, first introduced by the EULER genome assembly package. Unlike EULER and similar programs, Quake utilizes a robust mixture model of erroneous and genuine k-mer distributions to determine where errors are located. Then Quake uses read quality values and
learns the nucleotide to nucleotide error rates to determine what types of errors are most likely. This leads to more corrections and greater accuracy, especially with respect to avoiding mis-corrections, which create false sequence unsimilar to anything in the original genome sequence from which the read was taken.

QUality ASsesment Tool for Genome Assembly QUAST evaluates a quality of genome assemblies by computing various metrics and providing nice reports.

QuickTree is a program for the rapid reconstruction of phylogenies by the Neighbor-Joining method. For details, see the article published in the journal 'Bioinformatics' (18:1546-1547).

R is a language and environment for statistical computing and graphics. In the context of the analysis of genomic data, R includes some statistical packages for clustering, linear model, anova, ...(downloaded from the CRAN). There is also others packages dedicated for the microarray analysis (downloaded from the CRAN). The last the R-project about bioanalysis is named bioconductor (http://www.bioconductor.org/) for the analysis and comprehension of genomic data. The packages anapuce and varmixt developed by the team Statistique et génome (OMIP department INA P-G & INRA - http://www.inapg.fr/ens_rech/mathinfo/recherche/mathematique/outil.html) for differential analysis are also available on the platform.

Rainbow package consists of several programs used for RAD-seq related clustering and de novo assembly.

RATT is software to transfer annotation from a reference (annotated) genome to an unannotated query genome.

RAxML (Randomized Axelerated Maximum Likelihood) is a program for sequential and parallel Maximum Likelihood [1] based inference of large phylogenetic trees. It has originally been derived from fastDNAml which in turn was derived from Joe Felsentein’s dnaml which is part of the PHYLIP [2] package.

Ray is a parallel de novo genome assembler that utilises the message-passing interface everywhere and is implemented using peer-to-peer communication.

Read and reformat biosequences

ReAS: Recovery of Ancestral Sequences for Transposable Elements from the Unassembled Reads of a Whole Genome Shotgun

Recherche de TFBS et ECR dans des sequences homologues. Pas d’alignement necessaire en input, pas de prerequis de PWM. Mesure de la conservation relative entre les sequences par recherche d’oligo conserves et scoring de similarite globale entre deux sequences homologues. Permet de chercher aussi les enhancers distaux. Fonctionnerait sur des promoteurs non annotes (pas de TSS connu).

RepeatMasker

RepeatScout is a tool to discover repetitive substrings in DNA.
Reptile is a software developed in C++ for correcting sequencing errors in short reads from next-gen sequencing platforms.

RNAmmer 1.2 predicts 5s/8s, 16s/18s, and 23s/28s ribosomal RNA in full genome sequences.

RUM is an alignment, junction calling, and feature quantification pipeline specifically designed for Illumina RNA-Seq data.

SAM Tools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.

SAS - Statistical Analysis System

SCWRL3.0 is a completely new version of the SCWRL program for prediction of protein side-chain conformations. SCWRL3.0 is based on a new algorithm based on graph theory that solves the combinatorial problem in side-chain prediction more rapidly than any other available program. SCWRL3.0 is more accurate than previous versions of SCWRL, while the new algorithm will allow for development of more sophisticated energy functions and for incorporation of side-chain flexibility around rotameric positions.

sDIALIGN is a software program for multiple alignment developed by Burkhard Morgenstern et al. While standard alignment methods rely on comparing single residues and imposing gap penalties, DIALIGN constructs pairwise and multiple alignments by comparing whole segments of the sequences. No gap penalty is used. This approach is especially efficient where sequences are not globally related but share only local similarities, as is the case with genomic DNA and with many protein families.

SeaView is a graphical multiple sequence alignment editor. SeaView is able to read and write various alignment formats (NEXUS, MSF, CLUSTAL, FASTA, PHYLIP, MASE)

Seq-Gen is a program that will simulate the evolution of nucleotide or amino acid sequences along a phylogeny, using common models of the substitution process. A range of models of molecular evolution are implemented including the general reversible model. State frequencies and other parameters of the model may be given and site-specific rate heterogeneity may also be incorporated in a number of ways. Any number of trees may be read in and the program will produce any number of data sets for each tree. Thus large sets of replicate simulations can be easily created. It has been designed to be a general purpose simulator that incorporates most of the commonly used (and computationally tractable) models of molecular sequence evolution.

SeqMap is a tool for mapping large amount of oligonucleotide to the genome. It is designed for finding all the places in a genome where an oligonucleotide could potentially come from. SeqMap can efficiently map as many as dozens of millions of short sequences to a genome of several billions of nucleotides. While doing the mapping, several mutations as well as insertions/deletions of the nucleotide bases in the sequences can be tolerated and furthermore detected. Various input and output formats are supported, as well as many command line options for tuning almost every steps in the mapping process.

SeqUence Repository and Feature detectionsNucleotidic sequence production commonly involve several dedicated bioinformatic softwares for sequence basecalling, vector detection, etc.
SIM4 recherche les meilleurs alignements locaux entre une séquence d'ADNC et une séquence d'ADN génomique (ARNm, EST) contenant ce gène et autorisant la présence d'introns et un petit nombre d'erreurs de séquençage.

Similar to phrap, CAP3 takes individual sequences and assembles them into sequence.s

SimWalk2 is a statistical genetics computer application for haplotype, parametric linkage, non-parametric linkage (NPL), identity by descent (IBD) and mistyping analyses on any size of pedigree. SimWalk2 uses Markov chain Monte Carlo (MCMC) and simulated annealing algorithms to perform these multipoint analyses.

SNAP is a new sequence aligner that is 10-100x faster and simultaneously more accurate than existing tools like BWA, Bowtie2 and SOAP2. It runs on commodity x86 processors, and supports a rich error model that lets it cheaply match reads with more differences from the reference than other tools. This gives SNAP up to 2x lower error rates than existing tools and lets it match larger mutations that they may miss.

SOAPaligner/soap2 is a member of the SOAP (Short Oligonucleotide Analysis Package). It is an updated version of SOAP software for short oligonucleotide alignment. The new program features in super fast and accurate alignment for huge amounts of short reads generated by Illumina/Solexa Genome Analyzer. Compared to soap v1, it is one order of magnitude faster. It require only 2 minutes aligning one million single-end reads onto the human reference genome. Another remarkable improvement of SOAPaligner is that it now supports a wide range of the read length.

SOAPdenovo

SOAPdenovo is a novel short-read assembly method that can build a de novo draft assembly for the human-sized genomes. The program is specially designed to assemble Illumina GA short reads. It creates new opportunities for building reference sequences and carrying out accurate analyses of unexplored genomes in a cost effective way.

SolexaQA is a Perl-based software package that calculates quality statistics and creates visual representations of data quality from FASTQ files generated by Illumina second-generation sequencing technology (“Solexa”).

SPatt (Statistic for Patterns) is a suite of C++ programs designed for the computation of pattern occurrences p-value on text. Assuming the text is generated according to Markov model, the p-value of a given observation is its probability to occur. The lower is the p-value, the more unlikely is the observation. For example, this tools can be used to find patterns with unusual behaviour in DNA sequences.

SSAHA (Sequence Search and Alignment by Hashing Algorithm) is an algorithm for very fast matching and alignment of DNA sequences. It achieves its fast search speed by encoding sequence information in a perfect hash function.

SSAKE is a genomics application for assembling millions of very short DNA sequences. It is an easy-to-use, robust, reliable and tractable clustering algorithm for very short sequence reads, such as those generated by Illumina Ltd.

Stacks is a software pipeline for building loci out of a set of short-read sequenced samples. Stacks was developed for the purpose of building genetic maps from RAD-Tag Illumina sequence data, but can also be readily applied to population studies, and phylogeography.
STRIDE = Protein secondary structure assignment from atomic coordinates

Strides is a program to recognize secondary structural elements in proteins from their atomic coordinates.

SvcR is an implementation of an algorithm for clustering based on the search for a separator in a space of characteristics between points described in a space of data. The format of the data is defined by a table attribute/value (matrix). The data are transformed by a unique cluster defined by a ray of points with support. It can be used to define a ray of this ball in the space of data for reconstructing the frontier maintaining several clusters.

Tablet is a lightweight, high-performance graphical viewer for next generation sequence assemblies and alignments.

TagDust is a program to eliminate artifactual reads from next-generation sequencing data sets.

Tandem Repeats Finder

T-Coffee is a multiple sequence alignment package. Given a set of sequences (Proteins or DNA), T-Coffee generates a multiple sequence alignment. Version 2.00 and higher can mix sequences and structures.

TGI

TGI Clustering tools (TGICL): a software system for fast clustering of large EST datasets. This package automates clustering and assembly of a large EST/mRNA dataset. The clustering is performed by a slightly modified version of NCBI’s megablast, and the resulting clusters are then assembled using CAP3 assembly program. TGICL starts with a large multi-FASTA file (and an optional peer quality values file) and outputs the assembly files as produced by CAP3.

The `grepseq` program takes a keyword which can contain ambiguous characters and character classes (also called a fixed-width motif) and then searches files and databases for exact or approximate matches to that keyword. The program produces one of two kinds of output, either a list of the matching sequences with the places where the keyword matched, or the complete entries of sequences containing matches, where each entry is annotated with the places where the matches occur.

The 454 pyrosequencing reads contains artificially duplicates, which might lead to misleading conclusions. cd-hit-454 is a fast program to identify exact and nearly identical duplicates, the reads begin at the same position but may vary in length or bear mismatches. cd-hit-454 can process a dataset in ~10 minutes. it also provides a consensus sequence for each group of duplicates.

The ARB software is a graphically oriented package comprising various tools for sequence database handling and data analysis. A central database of processed (aligned) sequences and any type of additional data linked to the respective sequence entries is structured according to phylogeny or other user defined criteria.

The BEDTools utilities allow one to address common genomics tasks such as finding feature overlaps and computing coverage. The utilities are largely based on four widely-used file formats: BED, GFF/GTF, VCF, and SAM/BAM. Using BEDTools, one can develop sophisticated pipelines that answer complicated research questions by “streaming” several BEDTools together. The following are examples of common questions that one can address with BEDTools.
The core of the LINKAGE package is a series of programs for maximum likelihood estimation of recombination rates, calculation of lod score tables, and analysis of genetic risks.

The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing. Next-Generation sequencing machines usually produce FASTA or FASTQ files, containing multiple short-reads sequences (possibly with quality information).

The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.

The MCL algorithm is short for the Markov Cluster Algorithm, a fast and scalable unsupervised cluster algorithm for networks (also known as graphs) based on simulation of (stochastic) flow in graphs.

The program structure is a free software package for using multi-locus genotype data to investigate population structure. Its uses include inferring the presence of distinct populations, assigning individuals to populations, studying hybrid zones, identifying migrants and admixed individuals, and estimating population allele frequencies in situations where many individuals are migrants or admixed. It can be applied to most of the commonly-used genetic markers, including SNPs, microsatellites, RFLPs and AFLPs.

The recent version of Illumina's CASAVA pipeline (Version 1.8) produces FASTQ files with both reads that pass filtering and reads that don't. The new READ-ID (the @ line) contains many new fields, one of them indicates whether the read is filtered or not. This program can filter FASTQ files produced by CASAVA 1.8, and keep/discard reads based on this filter flag.

The SOLID System Analysis Pipeline Tool (Corona Lite) is an off-instrument SOLID data analysis software package. It supports functionality for mapping color space reads to large or small genomes, pairing for mate-pair runs, SNP calling and generating consensus sequences.

The Staden Package is a set of tools covering sequence assembly, editing and analysis. Gap4 performs sequence assembly, contig ordering based on read pair data, contig joining based on sequence comparisons, assembly checking, repeat searching, experiment suggestion, read pair analysis and contig editing. Pregap4 provides a graphical user interface to set up the processing required to prepare trace data for assembly or analysis. Trev is a rapid and flexible viewer and editor for ABI, ALF, SCF and ZTR trace files. Prefinish analyses partially completed sequence assemblies and suggests the most efficient set of experiments to help finish the project. Tracediff and hetscan automatically locate mutations by comparing trace data against reference traces. Spin analyses nucleotide sequences to find genes, restriction sites, motifs, etc. It can perform translations, find open reading frames, count codons, etc.

The Vienna RNA Package consists of a C code library and several stand-alone programs for the prediction and comparison of RNA secondary structures. RNA secondary structure prediction through energy minimization is the most used function in the package. We provide three kinds of dynamic programming algorithms for structure prediction: the minimum free energy algorithm of (Zuker & Stiegler 1981) which yields a single optimal structure, the partition function algorithm of (McCaskill 1990) which calculates base pair probabilities in the thermodynamic ensemble, and the suboptimal folding algorithm of (Wuchty et.al 1999) which generates all suboptimal structures within a given energy range of the optimal energy. For secondary structure comparison, the
package contains several measures of distance (dissimilarities) using either string alignment or tree-editing (Shapiro & Zhang 1990). Finally, we provide an algorithm to design sequences with a predefined structure (inverse folding).

tmhmm is one of the better prediction methods of transmembrane helices in proteins.

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

TREE-PUZZLE is a computer program to reconstruct phylogenetic trees from molecular sequence data by maximum likelihood. It implements a fast tree search algorithm, quartet puzzling, that allows analysis of large data sets and automatically assigns estimations of support to each internal branch. TREEPUZZLE also computes pairwise maximum likelihood distances as well as branch lengths for user specified trees. Branch lengths can be calculated with and without the molecular-clock assumption. In addition, TREE-PUZZLE o ers likelihood mapping, a method to investigate the support of a hypothesized internal branch without computing an overall tree and to visualize the phylogenetic content of a sequence alignment. TREE-PUZZLE also conducts a number of statistical tests on the data set (chi-square test for homogeneity of base composition, likelihood ratio to test the clock hypothesis, one and two-sided Kishino-Hasegawa test, Shimodaira-Hasegawa test, Expected Likelihood Weights). The models of substitution provided by TREE-PUZZLE are GTR, TN, HKY, F84, SH for nucleotides, Dayhoff, JTT, mtREV24, BLOSUM 62, VT, WAG for amino acids, and F81 for two-state data. Rate heterogeneity is modeled by a discrete Gamma distribution and by allowing invariable sites. The corresponding parameters (except for GTR) can be inferred from the data set.

TribeMCL is a method for clustering proteins into related groups, which are termed 'protein families'. This clustering is achieved by analysing similarity patterns between proteins in a given dataset, and using these patterns to assign proteins into related groups.

VCAKE is a genetic sequence assembler capable of assembling millions of small nucleotide reads even in the presence of sequencing error. This software is currently geared towards de novo assembly of Illumina’s Solexa Sequencing data.

Version améliorée du programe d'annotation de données métagénomiques Metagene. Prediction de genes procaryotes à partir d’un génome ou d'un set de génomes anonymes. Particulièrement adapté aux analyses métagénomiques.

Vmatch replaces Reputer. It looks for all possible repeats in genomes, withs possibility to specify the kind of repeats to look for, like its identitypercentage, minimal length, etc...Can also be used to mask repeats inssequences, to analyze repeat families, etc...

Washington University BLAST (WU BLAST) version 2.0 is a powerful software package for gene and protein identification, using sensitive, selective and rapid similarity searches of protein and nucleotide sequence databases.

Within EMBOSS you will find around 100 programs (applications). These are just some of the areas covered (Sequence alignment, Rapid database searching with sequence patterns,Protein motif identification, including domain analysis, Nucleotide sequence pattern analysis, for example to identify CpG islands or repeats, Codon usage analysis for small genomes, Rapid identification of
sequence patterns in large scale sequence sets, Presentation tools for publication...)

WOMBAT is a program to facilitate analyses fitting a linear, mixed model via restricted maximum likelihood (REML). It is assumed that traits analysed are continuous and have a multivariate normal distribution.

X-PLOR is a program system for computational structural biology. X-PLOR stands for exploration of conformational space of macromolecules restrained to regions allowed by combinations of empirical energy functions and experimental data. But it also stands for exploration of modern concepts of structured programming in macromolecular simulation.

YASS est un outil permettant la recherche locale de similarités dans les séquences d'ADN.