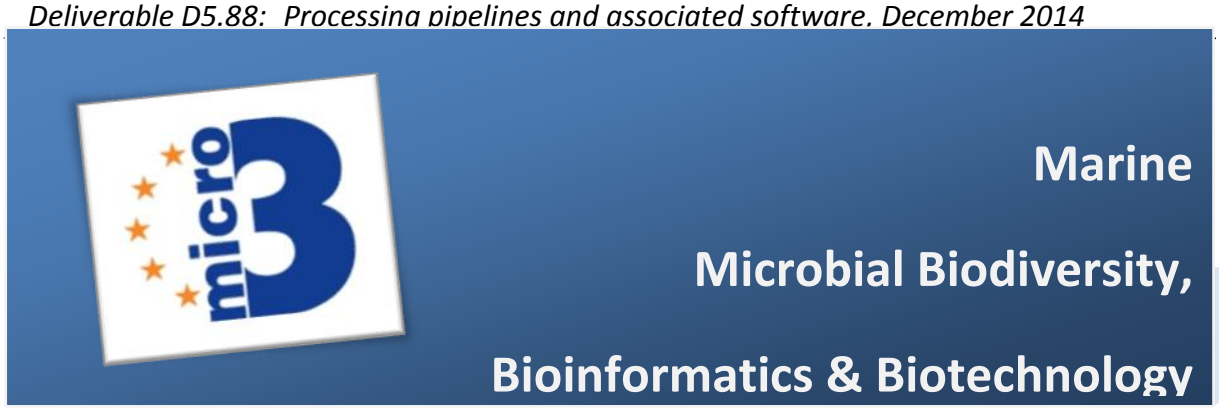

Grant agreement n°287589

Acronym: Micro B3

Start date of project: 01/01/2012, funded for 48 month

Deliverable 5.88

Processing pipelines and associated software – Protists annotation pipeline

Version: 1

Circulated to: Authors and Coordinator (05/01/2015)

Approved by: Frank Oliver Glöckner (17/01/2015)

Expected Submission Date: 31/12/2014

Actual Submission Date: 19/01/2015

Lead Party for Deliverable: Jaillon Olivier, Genoscope
Mail: ojaillon@genoscope.cns.fr

Public (PU)	X
Restricted to other programme participants (including the Commission Services) (PP)	
Restricted to a group specified by the consortium (including the Commission Services) (RE)	
Confidential, only for members of the consortium (including the Commission Services) (CO)	

Contents

Summary

Recent projects on marine plankton focus not only on the diversity of prokaryotes, but also address eukaryotic organisms in a global ecosystemic view. Most of these eukaryotic organisms are taxonomically far from sequenced genomes. In this context, obtaining sequences of their coding genes is technically challenging but it is necessary to have robust methods.

For this purpose, we developed an annotation pipeline able to structurally annotate any marine eukaryotic genome, especially protists. The core of the method is an integration of different complementary information, but is particularly suited to combine protein information (available in public databases for instance), ab initio prediction (with appropriate calibration) and metatranscriptomic data. This pipeline doesn't need specific RNA-seq data to completely and efficiently annotate protists sequences, permitting the annotation of uncultured organism genomes from large-scale projects samples.

Introduction

Marine eukaryotes are weakly represented in public genomic databases. This is especially true for planktonic species that are targeted in projects such as Tara Oceans. The paucity of knowledge on these organisms challenges gene prediction since standard methods rely on training on descriptions of known genes. Many parameters describing the exon/intron structure, lengths and number of exons and introns, frequencies of splicing sites, and even the genetic code can differ between species. Public availability of transcriptome and metatranscriptome sequence data is increasing. This might represent an excellent opportunity to detect coding sequences, and splicing information without any *a priori* knowledge.

However, this increase will also exacerbate CPU intensiveness and impact the computational feasibility of the pipeline.

We developed a procedure to detect genes on a given eukaryote genome sequence taking in account the following constraints:

- Existence of minimal (or no) *a priori* information
- Maximum automation
- Optimization of CPU time, adaptable on each query depending on associate resources
- Output of a set of genes in a standard format
- Production of a final gene model that integrates number and types of evidence that are variable according to genomes

Pipeline Presentation

We designed an annotation pipeline combining 3 types of resources to maximize the number of detected genes (Figure 1):

- Protein alignments. This resource is useful to annotate conserved genes with relative confidence. However, for distant organisms only a small fraction of genes will have – partial – matches.
- *ab initio* predictions, without the need of an *a priori* knowledge: only a sample of Open Reading Frames (ORFs) are necessary to calibrate this method. Moreover, it produces complete models, from start to stop.
- Metatranscriptomics. As the number of metatranscriptomic projects increases, it is interesting to retrieve specific data into the huge collections they generate.

Gene models from these three resources are then combined together to maximize the number and the completion of detected genes.

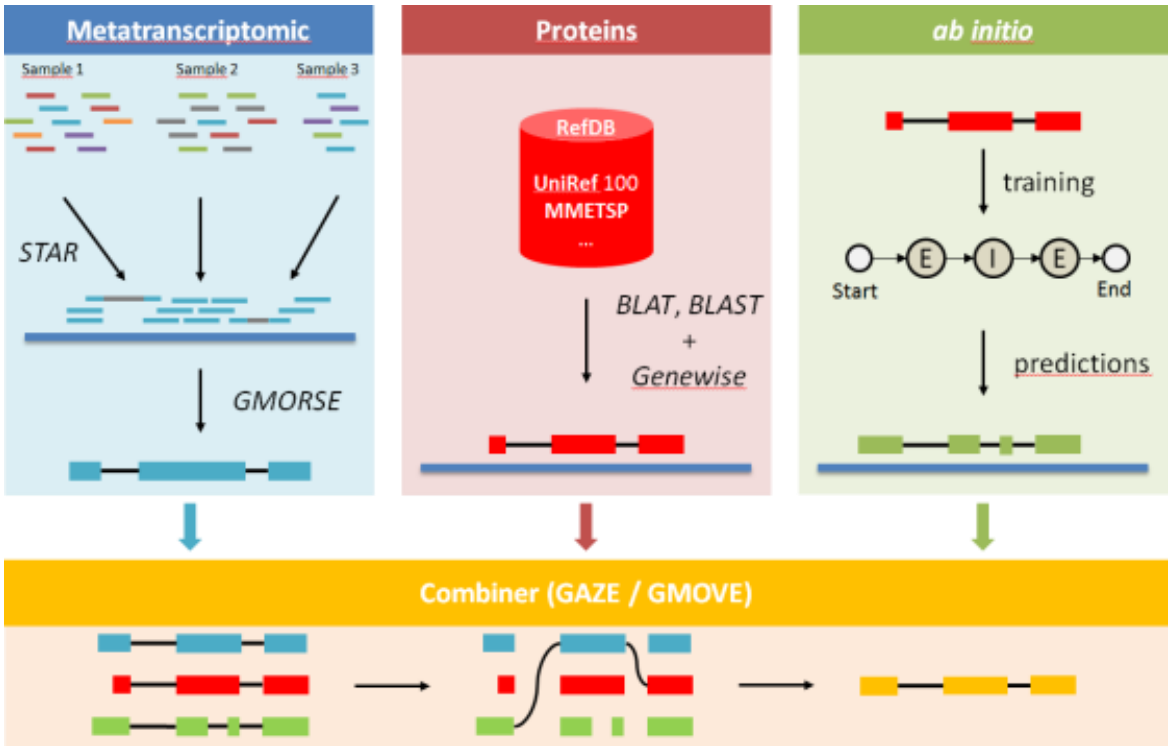

Figure 1: Annotation pipeline. 3 types of resources are used: 1 – metatranscritpomic data (blue) from Tara Oceans samples. Reads are aligned to the assembly and GMORSE produces gene models from reads coverage. 2 – Proteins from a custom marine eukaryotic database. 3 – *ab initio* prediction using SNAP program, trained on protein matches. All the results are combined with the GMOVE tool to output gene models.

In order to avoid problems related to low complexity regions or repetitive elements, a masking step is first performed on the genomic sequences. Indeed, low complexity regions can recruit non specific reads from metatranscriptomic collections or non specific proteins from databases, and so lead to false positive gene models.

Repetitive elements are often transposable elements [1] in eukaryotes, and need a specific method to be annotated.

A – Masking

The masking step uses RepeatMasker [2] and RepeatScout [3] tools. The former removes low complexity regions in order to avoid non specific mapping and thus reduce mapping computational time. RepeatScout is used to mask more complicated repeats such as transposable elements which cannot be correctly annotated without a specific pipeline.

Masking is a necessary step to save computational time and to avoid false positive gene annotations.

For example, 5 to 10% of the genome of marine stramenopiles such as Chrysophytes or MAST is masked using this procedure. This proportion is consistent with relatively small eukaryotic genomes (estimated to 40-50 Mbp) with few transposable elements.

B – Protein alignments

We designed the pipeline to use a large collection of proteins: for example, we built a custom database of proteins grouping Uniref100 [4], MMETSP [5] and non public transcriptomes of marine eukaryotes. As this database contains more than 33 million proteins, optimizations are needed to compute alignments of this dataset in a reasonable amount of time. This is why a multi-step process is in place.

The first step is a low sensitive mapping using BLAT [6]. The 10 best matches are then correctly aligned with GeneWise [7] to properly define exons and introns. We limit the maximum number of matches in order to significantly improve CPU and memory consumption for conserved proteins: 10 proteins at most are aligned with GeneWise (more CPU intensive than BLAT) on the same region instead of hundreds or thousands for some cases. Sensitivity is not impacted since the most informative alignments are kept.

The second step is a more sensitive mapping with BLAST (blastx) [8] on non matching regions from previous BLAT step. GeneWise is then used on the first 10 best matches similarly to the first step. The BLAST step is very important considering unknown organisms not represented in databases (Table 1)

Method	# Matches	# loci
BLAT + GeneWise	16,235	1,619
BLAST + GeneWise	12,510	2,422
Total	**28,745**	**4,041**

Table 1: Result of the 2 pass protein alignment on a marine Chrysophyte Clade C. The second line demonstrates the importance of a sensitive comparison to align distant proteins on unknown organisms.

At each stage, the database is split in independent smaller packages to parallelize computations on a computer cluster. Another option to reduce computational time would be the use of a less redundant database, based on UniRef90 for example.

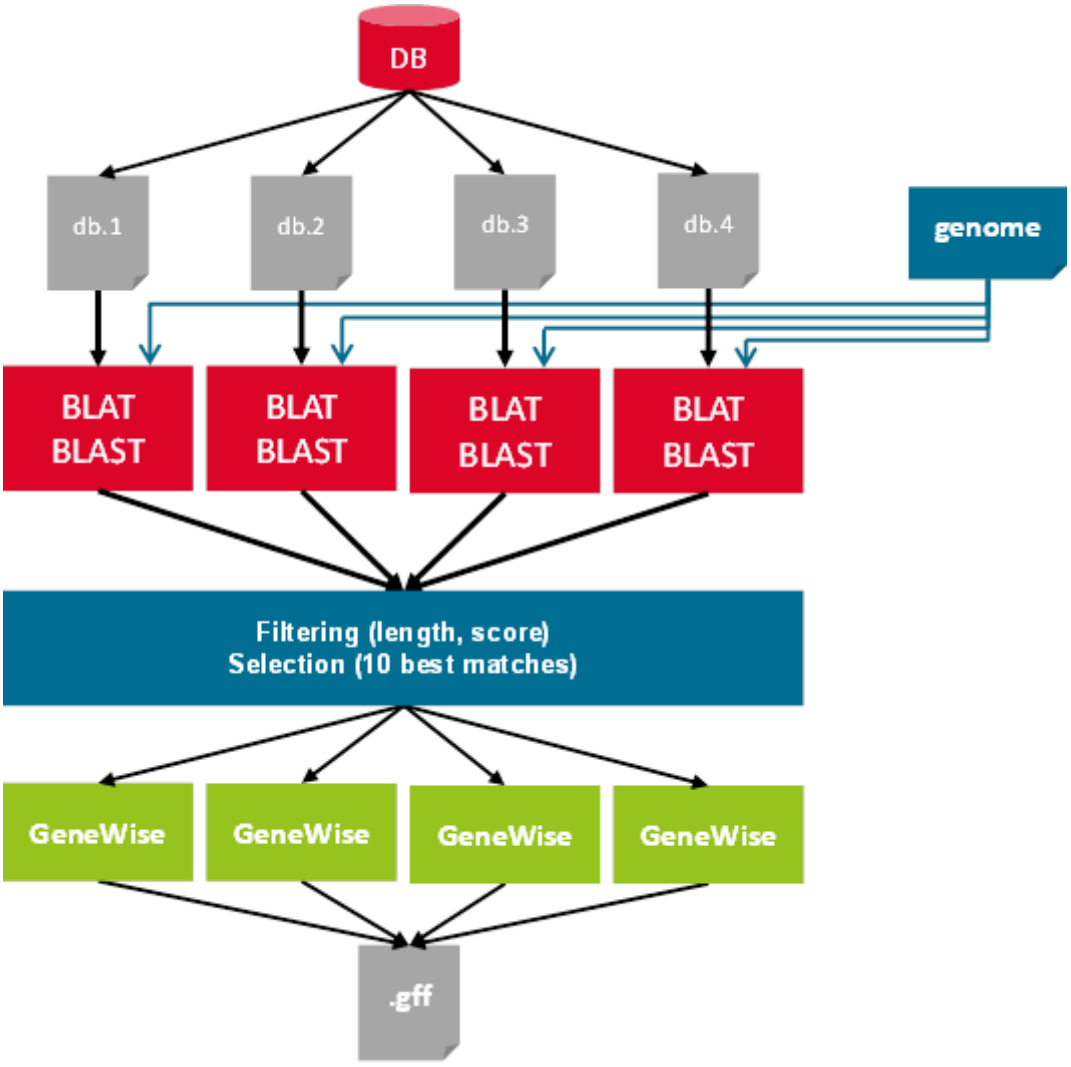

Figure 2: Parallelization of the protein mapping pipeline. The database (DB) is split into independent packages (db.x). Each package is aligned to the genome using BLAT (pass 1) or BLAST (pass 2). For each locus, the 10 best matches are selected, based on score (BLAT) or e-value (BLAST), and are aligned with GeneWise to the genome.

C - *ab initio* gene predictions

We use SNAP [9] as eukaryotic gene predictor. Based on a Markov Model, it needs a calibration on hundreds of coding sequences before running on all the assembly. We calibrate it with information from protein matches or GMORSE models.

	# Genes	Exons / Gene	# True positives	# False negatives
Reference Annotation	7,807	1.14		
SNAP predictions	5,815	2.12	5,523 (95%)	2,284 (29%)

Table 2: Overlapping between SNAP predictions and reference annotation [10] of Bathycoccus *prasinos* RCC1105. SNAP misses 2,284 genes but 95% of all predicted genes overlap at least 10% of a gene from the official annotation.

The main advantage of *ab initio* predictions is the fact that it outputs complete models, from start to stop. SNAP models can then extend incomplete gene models generated from biological evidences.

D - Metatranscriptomics

Tara Oceans project will release several large metatranscriptomic datasets, with a total of approximately 100 billion illumina paired end reads. Mapping all these reads on a eukaryotic genome is very challenging: aligners have to deal with complexity of aligning short reads and non exact mapping since they have to allow gaps (introns).

We developed a two-step approach: first, we operate a fast selection of reads using kfir program (not published) based on k-mer sharing. Then, we align this subset with STAR splice aligner [11].

Reads selection

Kfir – K-mer filtering of Reads – is used to obtain a subset of reads matching a reference genome. It consists of two major steps: the dictionary creation and the validation of reads.

Dictionary creation: Each sequence of the genome is split into sliding k-mers (default k size is 25 bp, proved to be specific enough without losing sensitivity) and an integer is associated with each forward and reverse k-mer, of which only one is conserved. Integers associated with k-mers are stored in a hash structure, to allow optimal access to the data during the second phase.

Validation of reads: either single or synchronized paired-end read files may be used as input. Each read is, as the genomic reference, split into k-mers to be tested against the word dictionary. A notable amelioration to the program has been the addition of sequence complexity test to be sure to retrieve only informative reads.

Alignments

After selection, candidates are aligned using STAR aligner. Default parameters are used, but a maximum intron size is often defined because STAR tends to create artificial gaps to align reads perfectly. This leads to 'gaps' (considered as introns) about the size of the scaffold.

The output bam is parsed to filter out alignments with less than 95% identity on all the alignment (except in introns) and low complexity sequences (using DUST program).

This crucial step can be a bottleneck for computations, depending on the number of reads selected by kfir: the more abundant is a genome in the metatranscriptomic data, the more time it will need to perform alignments, but the more models we will generate.

Models generation

The GMORSE tool [12] then creates gene models from coverage information. At the moment, GMORSE needs a threshold coverage to create covtigs (coverage contigs, potential exons) and junctions (potential introns), so a close look at the coverage distribution is necessary before running this step.

GMORSE models with coding sequence shorter than 90 bp are discarded, and we then keep models with the longest coding sequence for overlapping models.

E - Reconciliation

The reconciliation step was formerly done using GAZE [13]. But it appears that this tool is not well suited when we have partial evidences and we noted some strange decisions in structure choices. We have recently begun to use GMOVE (Gene Modelling using Various Evidences, not published yet), a combiner developed at *Genoscope*, relying more on resources than GAZE.

GMOVE is based on a graph structure where each vertex represents an exon and each edge an intron, as given by the input data. A path in the graph is then a transcript model. Several paths on the same genomic locus are representative of alternate transcripts.

GMOVE looks for an Open Reading Frame in every path and may extend the boundaries of the model to find a Methionine and/or a Stop codon and thus output the most complete model.

The models are the combination of different resources but are solely based on biological evidences.

Finally, we select the best models at each locus regarding several criteria:

- ORF length: we select models with the longest ORF.
- Intron size: models with very large intron size are discarded. Threshold depends on intron size distribution.
- Number of high confidence introns: confidence is evaluated from the ‘coverage’ in gaps from aligned spliced reads.

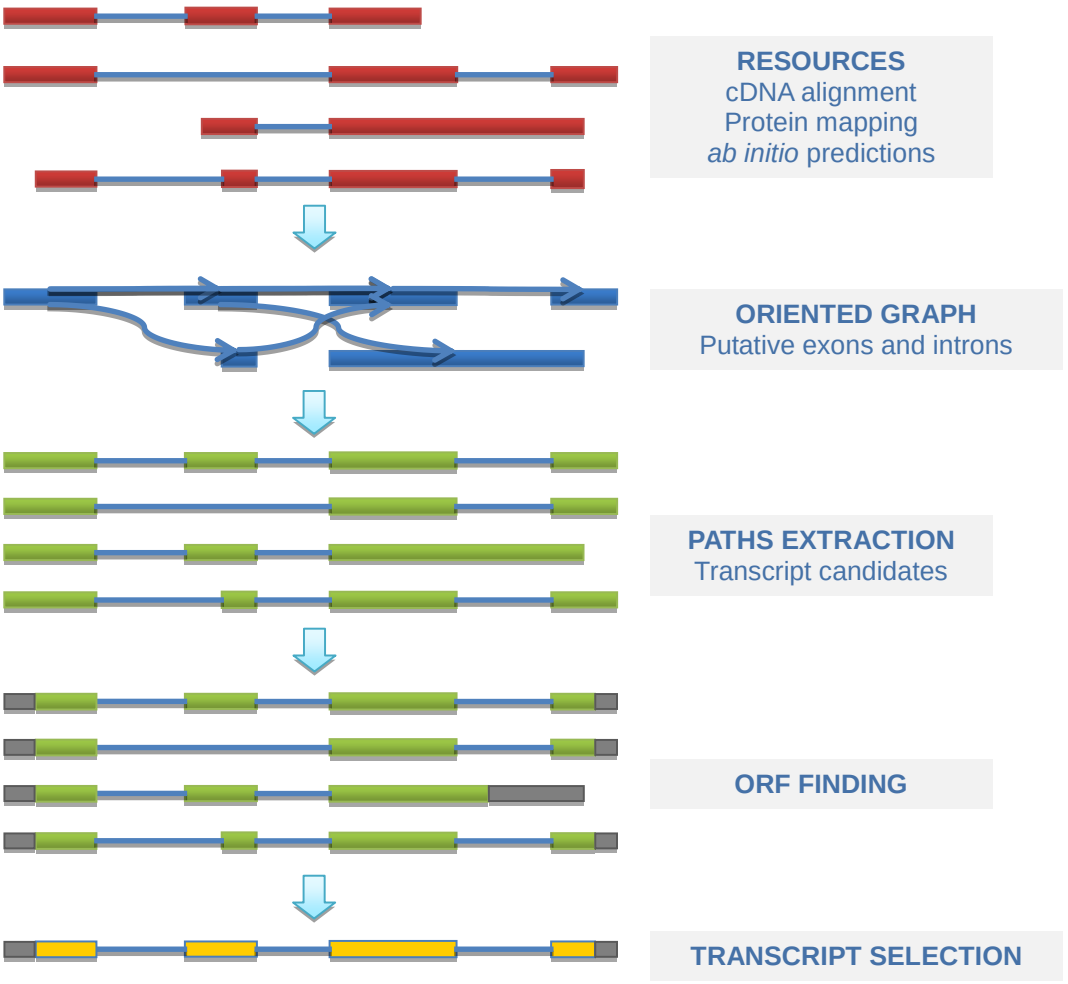

Figure 3: GMOVE algorithm. An oriented graph is built from all resources where exons (nodes) are linked by introns (edges). All possible paths are potential gene models. A filtering step selects the best models according to ORF length.

The output is a GFF [12] formatted annotation, a standard format allowing an easy loading into a genome browser such as a Generic Genome Browser (Figure 6).

Conclusions

Here we described a pipeline that allows us to annotate marine eukaryotic genomes without using RNAseq experiments. Instead, it uses metatranscriptomics data from Tara Oceans samples to retrieve specific transcriptomic reads. This original resource is then combined with more classical resources: protein alignments and *ab initio* predictions.

Several optimizations have been performed to reduce execution time:

- a subset of metatranscriptomics reads is selected before splice alignment
- proteins are aligned iteratively from the less sensitive method to the most sensitive one
- parallel computing is intensively used: during protein alignment, reads alignment and during reconciliation

This pipeline has been applied successfully on a dozen of marine uncultivated protists (mainly stramenopiles: MASTs and Chrysophytes) from Indian Ocean and Mediterranean. It will be used to annotate about a hundred of other protists from uncultivated phyla.

Protists pipeline References

[1] Wojciech Makałowski *et al*. Tranposable Elements and Their Identification. Evolutionary Genomics: Statistical and Computational Methods, Volume 1. Methods in Molecular Biology, vol. 855, p. 337-359.

[2] Smit, AFA, Hubley, R. (2008) RepeatModeler. *Open-1.0*. 2008-2010.

[3] Price A.L., Jones N.C. and Pevzner P.A. (2005). De novo identification of repeat families in large genomes. Proceedings of the 13 Annual International conference on Intelligent Systems for Molecular Biology (ISMB-05). Detroit, Michigan.

[4] Baris E. Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder and Cathy H. Wu. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics (2007) 23 (10): 1282-1288*

[5] Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, et al. (2014) The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol* 12(6): e1001889. doi:10.1371/journal.pbio.1001889

[6] Kent WJ. (2002) BLAT - the BLAST-like alignment tool. *Genome Res*. 2002 Apr;12(4):656-64.

[7] Birney E, Clamp M, Durbin R. (2004) GeneWise and Genomewise. *Genome Res. 2004;14:988–995. doi: 10.1101/gr.1865504.*

[8] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol. 215:403-410.*

[9] Ian Korf (2004) Gene finding in novel genomes. *BMC Bioinformatics 5:59*

[10] Moreau H, Verhelst B, Couloux A, Derelle E, Rombauts S, et al. Gene functionalities and genome structure in *Bathycoccus prasinos* reflect cellular spe5cializations at the base of the green lineage. *Genome Biology* doi:10.1186/gb-2012-13-8-r74

[11] A. Dobin et al (2012) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* doi: 10.1093/bioinformatics/bts635

[12] Denoeud, F. et al. (2008) Annotating genomes with massive-scale RNA sequencing. Genome Biol. 9, R175.

[13] Kevin L. Howe, Tom Chothia, Richard Durbin, GAZE: A Generic Framework for the Integration of Gene-Prediction Data by Dynamic Programming *Genome Res.*, Vol. 12, No. 9. pp. 1418-1427, doi:10.1101/gr.149502