SynBlast: assisting the analysis of conserved synteny information

Jörg Lehmann[1], Peter F. Stadler[1,2,3,4,5], and Sonja J. Prohaska[1,4,5]

[1]Bioinformatics Group, Department of Computer Science, University of Leipzig, Germany

[2]Interdisciplinary Center for Bioinformatics, University of Leipzig, Germany [3]Fraunhofer Institute for Cell Therapy and Immunology, Leipzig, Germany [4]Institute for Theoretical Chemistry, University of Vienna, Austria [5]Santa Fe Institute, Santa Fe, New Mexico, USA

Email: joe@bioinf.uni-leipzig.de **WorldWideWeb:** http://www.bioinf.uni-leipzig.de/

Motivation

One of the key tasks in comparative genomics is the retrieval of evolutionarily related genes within and between sequenced genomes. Fully-automated ortholog detection methods often fail to identify individual members in cases of large gene families, or to distinguish missing data from traceable gene losses. This situation can be improved in many cases by including conserved synteny information.

duplication, gene loss, different evolutionary rates, convergent evolution $\longrightarrow$ **pairwise sequence comparisons alone are not sufficient for ortholog search**

synteny: physical co-localization of genetic loci on a chromosome

conserved synteny: local maintenance of gene content and order

rearrangements: e.g. translocations, inversions, ...

orthologs: a pair of genetic loci (different species) that arose from each other through speciation (often also functionally preserved)

paralogs: a pair of genetic loci (different or same species) that arose from each other through duplication

duplication: small-scale/local, large-scale/non-local or retrotransposition

gene loss: nonfunctionalization/pseudogenization

Conserved synteny - an indicator for orthology?

Valuable practical information through conserved synteny information for the analysis of gene families:

- a known orthology relationship of a pair can be utilized to more reliably infer the relationships between its neighbors, respectively, as **linked genes likely share their duplication history**
- extensive conserved synteny often reflects functional constraints of syntenic genes, e.g. shared regulatory mechanisms
- conserved synteny even provides direct evidence for or against orthology if the genomic context of duplicate genes has sufficiently diverged prior to a speciation.
- **gene loss becomes traceable and** better **distinguishable from missing data**

SynBlast - another orthology pipeline?

Intentions of SynBlast

- **assist a manual curation of high-quality orthology assignments** based on complete target genome sequence data instead of pre-determined proteome sets
- present plausible homologs and their genomic context ranked according to different criteria for orthology including conserved synteny
- facilitate manual analysis in case of conflicting information of orthology (e.g. highly diverged genes or high rates of gene loss)

Main features

- command line Perl scripts for the pipeline steps
- uses Ensembl Core and Compara databases for reference retrieval and comparison (via Ensembl API)
- offers graphical representations of homologous regions from the reference cluster's perspective (dot-plots)
- **no automatized decision for/against orthology, but plausible rankings** of blast hit regions and further information (similarity score of HSP chains, global rank, intra-inter-score ratio, query-coverage, ...)

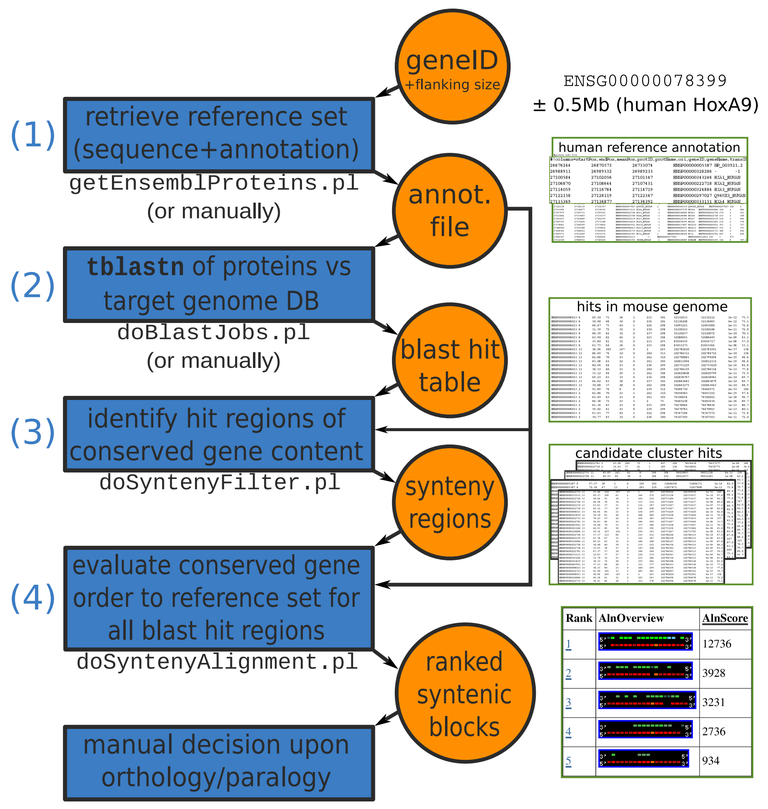

Figure 1: SynBlast "semi-automatic" pipeline implemented as a suite of Perl scripts.

Extraction and evaluation of candidate regions

1. extraction of relevant regions of target genome based on blast output - doSyntenyFilter.pl
   - divides blast hits into syntenic subsets restricted in its max. genomic range and requiring a minimum number of detected reference proteins (synteny regions)
2. evaluation of each of these subsets via gene order alignments to the reference region - doSyntenyAlignment.pl
   - the relative order of reference genes and their corresponding tblastn-hits (HSP chains) is compared via a two-level alignment procedure utilizing the blast similarity bit-score as match score of alignment characters

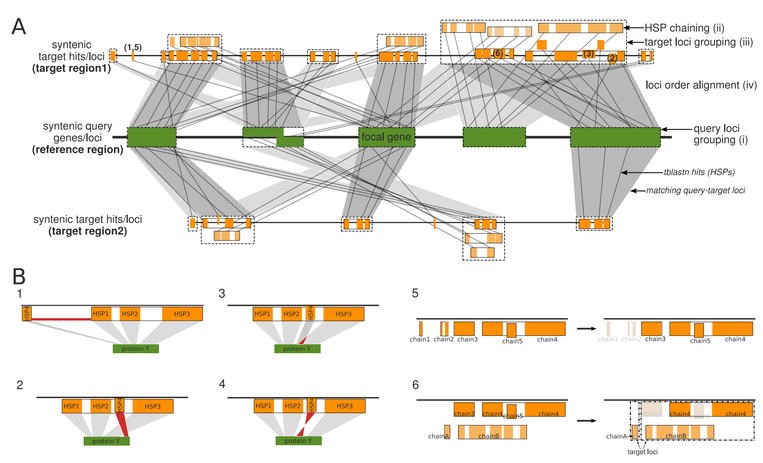

Figure 2: Evaluation of syntenic target regions.

- for each protein its HSPs are chained into approximate gene models (HSP chains) on the target region
- overlapping HSP chains are combined into target loci
- the relative order of target loci and ref. proteins is aligned to obtain a *gene order alignment* score
- additionally, the *(log)RatioSum* score is calculated to better discriminate ortholog clusters from paralogs
- candidate regions are ranked according to scores and visualized (dot-plot and its optimal gene order alignment), see Figure 3.

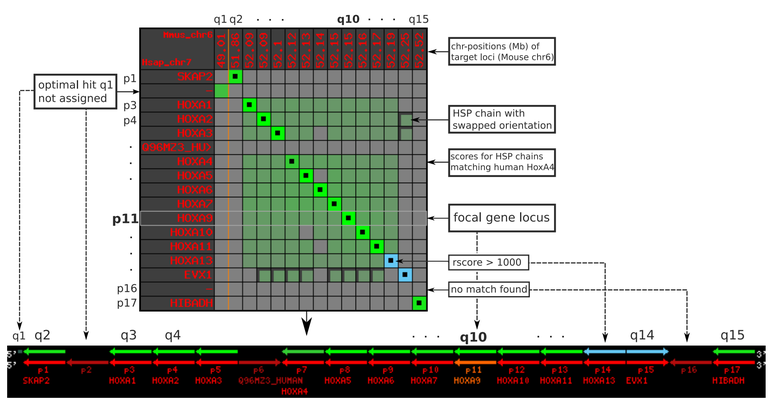

Figure 3: Dotplot graphic (top) and corresp. alignment graphic (human *HoxA* vs mouse). The optimal gene order alignment is represented by the set of dotted squares (optimal alignment path). Rows and columns of the scoring matrix correspond to linearly ordered query and target loci, respectively.

Application to *Hox*/*ParaHox* clusters

We applied SynBlast to two well-studied gene clusters, the vertebrate *Hox* and *ParaHox* clusters.

For the *Hox* cluster, using all human paralog clusters as references, cluster identities of vertebrate homologous clusters could clearly be assigned based on the SynBlast ranking, see Table 1 for an example.

	DrAa	*DrAb*	*DrBa*	*DrBb*	*DrCa1*	*DrCa2*	*DrCb*	*DrDa*
	chr.19	chr.16	chr.3	chr.12	chr.23	chr.23	chr.11	chr.9
Reference	10.5M	16M	23M	26.5M	33.7M	35.2M	0.6M	2M
HsA	**2.58**	**3.29**	-0.76	-0.39	-1.98	-1.98	-0.4	-0.74
HOXA9	**5581**	**4073**	4690	2119	3322	3318	1490	3269
chr.7	**2/1**	**1/3**	7/2	3/7	9/4	10/5	4/8	6/6
HsB	-1.14	-0.01	**3.13**	**0.42**	-1.66	-1.69	-0.64	-0.48
HOXB9	2702	1342	**6201**	**1647**	3008	2982	1003	1678
chr.17	6/4	3/7	**1/1**	**2/6**	7/2	8/3	5/8	4/5
HsC	-0.89	-1.51	-0.26	-0.34	**6.44**	**7.58**	**0.22**	-2.51
HOXC9	2361	2323	3108	1346	**8687**	**6537**	**4150**	3798
chr.12	7/6	8/7	5/5	6/8	**2/1**	**1/2**	**3/3**	9/4
HsD	-0.92	-0.63	-0.85	-0.6	-0.41	-0.41	-0.71	**1.76**
HOXD9	2799	1811	2660	871	3017	3013	1303	**4326**
chr.2	8/4	5/6	7/5	4/8	3/2	2/3	6/7	**1/1**

Table 1: SynBlast results for the 7 *Hox* clusters in *Danio rerio* (Zv7). The *logRatioSum* score, the gene order alignment score, and the corresponding ranks are given. Putative orthologs are depicted in bold. In combination, the two scores provide the best means to rank orthologous loci at the top. Note the artifactually duplicated *DrCa* cluster of chr.23.

With the help of conserved synteny information, we could suggest a plausible hypothesis [1] for the cluster identities of the highly fragmented *ParaHox* clusters in the zebrafish genome. The obviously *ParaHoxA* cluster regions of Danio are depicted in Figure 4 demonstrating the loss of the *Cdx2* gene for both regions.

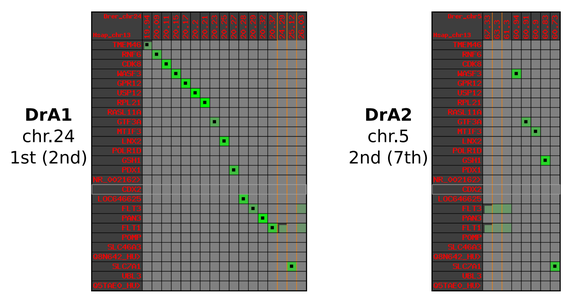

Figure 4: Dot-plots for the two high-ranking hits of the human *ParaHoxA* query region against the zebrafish genome. The DrA1 copy retained 13 of 24 genes flanking **the *Cdx2* locus** which **itself was obviously lost**. Here, the gene loss can reliably be distinguished from missing data due to well-conserved synteny information.

Terms and definitions

ratio of intra- and inter-score: quantifies the quality of an inter-species (potentially orthologous) hit t_s in relation to the similarity between closest-paralogs (q_{s_1}, q_{s_2}) in the reference genome $\longrightarrow$ measures the confidence in orthology:

$$S_{\text{intra}}(s)/S_{\text{inter}}(t_s) = \frac{b(s,q_{s_1}) - b(s,q_{s_2})}{b(s,q_{s_1})} \times \frac{b(s,q_{s_1})}{b(s,q_{s_1}) - b(s,t_s)} = \frac{b(s,q_{s_1}) - b(s,q_{s_2})}{b(s,q_{s_1}) - b(s,t_s)} \quad (1)$$

***(log)RatioSum* score:** defined as the sum of the (logarithms of) intra-inter-score ratios of all target loci assignments from gene order alignment

HSP: High scoring pair (single blast hit)

HSP chain: a maximal group of consistent HSPs for a query protein (its score is the sum of its HSPs' bit-scores)

target locus: set of HSP chains of different query overlapping on target coordinates

query locus: set of query proteins overlapping on reference coordinates

Availability

The SynBlast tool and tutorial as well as the Supplemental Material for our recent publication [1] is available at:
http://www.bioinf.uni-leipzig.de/Software/SynBlast/

References

[1] J. Lehmann, P. F. Stadler, and S. J. Prohaska. SynBlast: assisting the analysis of conserved synteny information. *BMC Bioinformatics*, 9:351, 2008.

Printed at Universitätsrechenzentrum Leipzig