A Workflow for Retrieving Orthologous Promoters and Implications for Workflow Management Systems. A Case Study.

Part

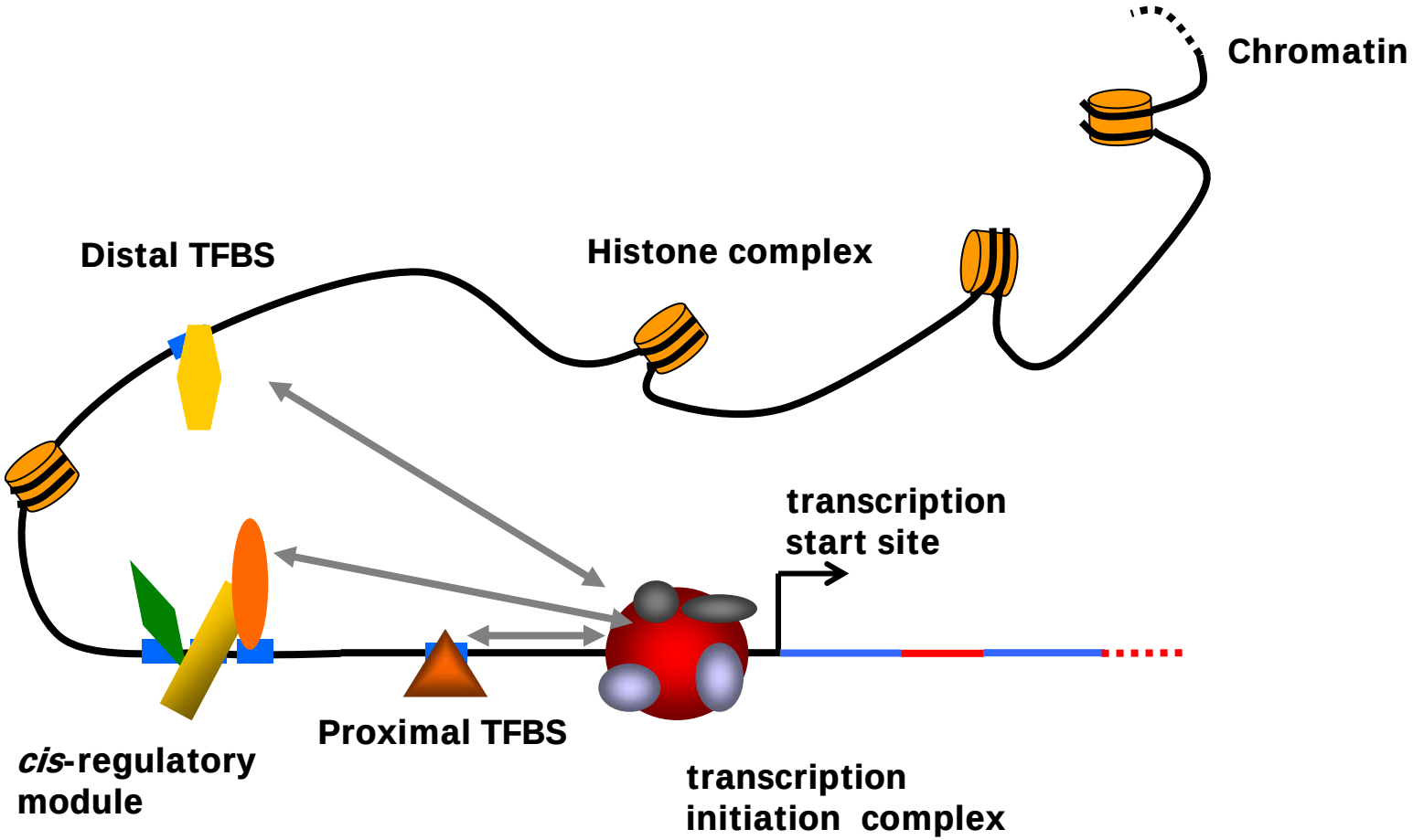
From Components to Processes in Bioinformatics

Department of Bioinformatics
Medical Faculty
Georg-August-University
Göttingen

Martin.Haubrock@bioinf.med.uni-goettingen.de

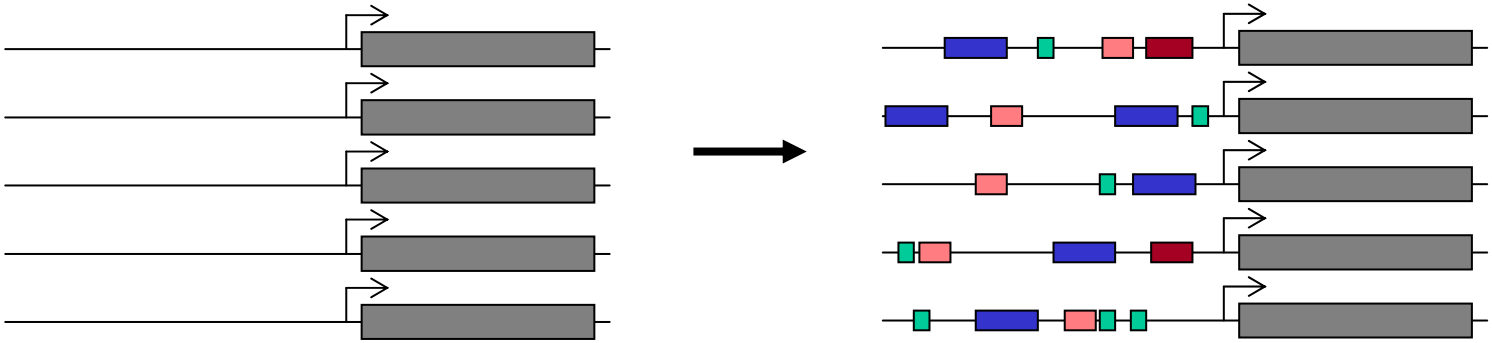
Components of transcriptional regulation

- **Transcription factors (TFs) bind to specific sites (transcription factor binding sites, TFBS) that are either proximal or distal to a transcription start site (TSS).**



Analysis of gene expression data

- **Promoter analysis of co-expressed genes**
  - Model:
    - Co-expression ~ Co-regulation
  - Given:
    - Set of potentially co-regulated genes
  - Task:
    - Find out the most likely set of transcription factor binding sites which could explain their co-regulation



Phylogenetic Footprinting

- **Prediction of potential TFBS using phylogenetic footprinting approach**

- **Idea:**
  - Not just coding regions, but also regulatory motifs are under a higher selective pressure than non-functional sections of a genome
  - Sequence alignments of regulatory regions can be used to identify potential conserved motifs between species.
  - A shared motif between many different species is assumed to more likely represent a real TFBS than a motif which is found in only one or a few species
  - We have developed a Hidden Markov Model which predicts potential TFBS using sequence alignments of regulatory regions and matrix representation of known TFs

Challenges in promoter retrieval

- **A unique and exact definition of a gene's promoter is a challenging task in computational biology:**
  - The majority of regulatory motifs are located within the -500 to -1 region upstream of a gene's transcribed region
  - In-silico gene prediction is still a challenging task in computational genomics
  - Experimental high-quality data on transcript start is very sparse
  - The predicted transcript start locations annotated in the common public genome databases are prone to be erroneous and cannot be taken for granted

Ensembl: human entity of the IL-2 gene

- **Genomic enviroment of the human IL-2 gene first exon:**
  - located on chromosome 4
  - 4 exons, 3 introns
  - transcript length: 1,044 bps
  - length of the first exon: 441 bps, ~300 bps untranslated

									tcaacaaatctaaacatttattcttttcatctgtttacttttgctcttgtccaccacaat atgctattcacatgttcagtgtagttttatgacaaagaaaatttttctgagttacttttgt atccccacccccttaaagaaaggaggaaaaactgtttcatacagaaggcgttaattgcat
1	ENSE00001293064	4	-1	123,596,899	123,597,339	-	0	441	GAATTAGAGCTATCACCTAAGTGTGGGCTAATGTAACAAAGAGGGATTTCACCTACATCC ATTCAGTCAGTCTTTGGGGGTTTAAAGAAATTCCAAAGAGTCATCAGAAGAGGAAAAATG AAGGTAATGTTTTTTCAGACAGGTAAAGTCTTTGAAAATATGTGTAATATGTAAAACATT TTGACACCCCCATAATATTTTTCCAGAATTAACAGTATAAATTGCATCTCTTGTTCAAGA GTTCCCTATCACTCTCTTTAATCACTACTCACAGTAACCTCAACTCCTGCCACAATGTAC AGGATGCAACTCCTGTCTTGCATTGCACTAAGTCTTGCACTTGTCACAAACAGTGCACCT ACTTCAAGTTCTACAAAGAAAACACAGCTACAACTGGAGCATTTACTGCTGGATTTACAG ATGATTTTGAATGGAATTAAT
	Intron 1-2	4	-1	123,596,809	123,596,898			90	gtaagtatatttccttttcttactaa.....ataacaatgcattatactttcttag
2	ENSE00000935280	4	-1	123,596,749	123,596,808	0	0	60	AATTACAAGAATCCCAAACTCACCAGGATGCTCACATTTAAGTTTTACATGCCCAAGAAG
	Intron 2-3	4	-1	123,594,459	123,596,748			2,290	gtaagtacaatattttatgttcaat.....gagctgatgataattattattctag
3	ENSE00000935278	4	-1	123,594,315	123,594,458	0	0	144	GCCACAGAACTGAAACATCTTCAGTGTCTAGAAGAAGAACTCAAACCTCTGGAGGAAGTG CTAAATTTAGCTCAAAGCAAAAACTTTCACTTAAGACCCAGGGACTTAATCAGCAATATC AACGTAATAGTTCTGGAACTAAAG
	Intron 3-4	4	-1	123,592,468	123,594,314			1,847	gtaaggcattactttatttgctctc.....aaaaattaacatttttcttttatag
4	ENSE00001138256	4	-1	123,592,080	123,592,467	0	-	388	GGATCTGAAACAACATTCATGTGTGAATATGCTGATGAGACAGCAACCATTGTAGAATTT

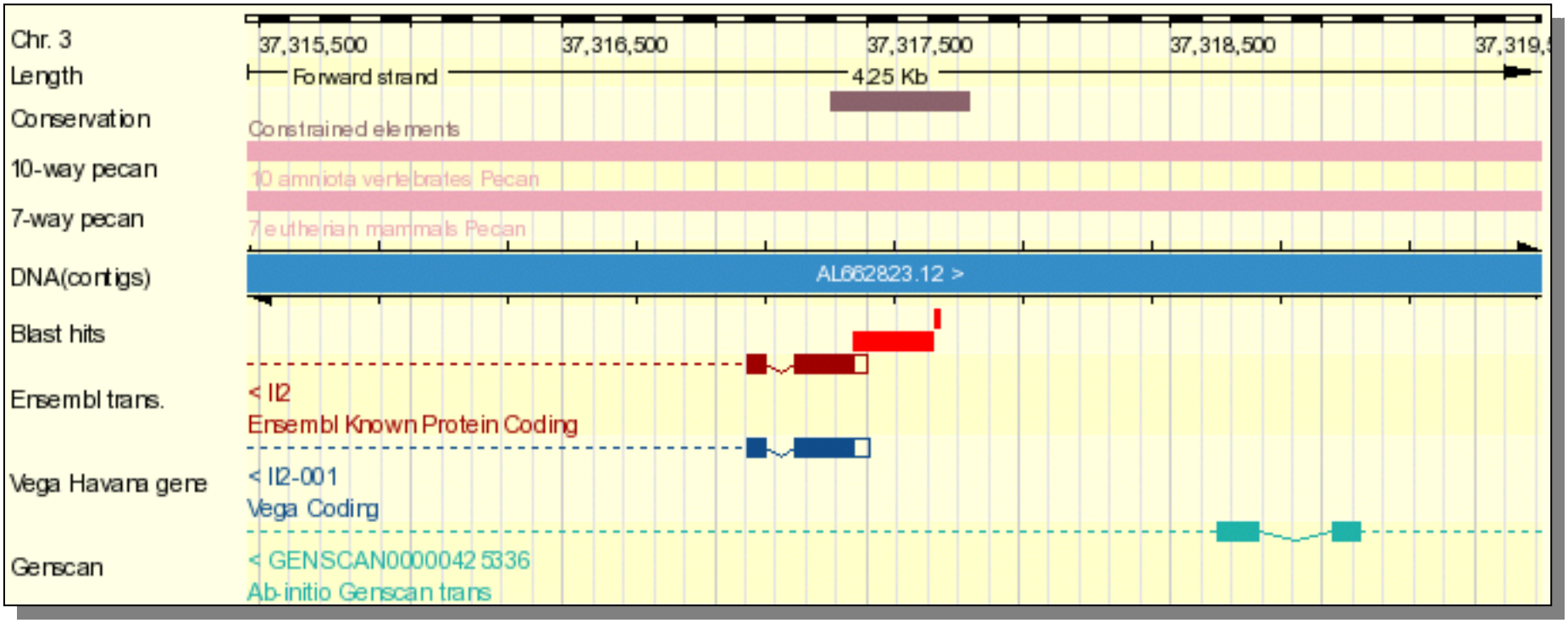
Ensembl: murine instance of the IL-2 gene

- **Genomic environment of the mouse IL-2 gene's first exon:**
  - located on chromosome 3
  - 3 exons, 2 introns
  - transcript length: 527 bps
  - length of first exon: 236 bps, ~50 bps untranslated

									gtgtatgggggtttaaagaaattccagagagtcatcagaagaggaaaaacaaaggtaatg ctttctgccacacaggtagactctttgaaaatatgtgtaatatgtaaaacatcgtgacac ccccatattatttttccagcattaacagtataaattgcctcccatgctgaagagctgcct
1	ENSMUSE00000345573	3	-1	37,317,267	37,317,502	-	0	236	ATCACCCTTGCTAATCACTCCTCACAGTGACCTCAAGTCCTGCAGGCATGTACAGCATGC AGCTCGCATCCTGTGTCACATTGACACTTGTGCTCCTTGTCAACAGCGCACCCACTTCAA GCTCCACTTCAAGCTCTACAGCGGAAGCACAGCAGCAGCAGCAGCAGCAGCAGCAGCAGC AGCAGCACCTGGAGCAGCTGTTGATGGACCTACAGGAGCTCCTGAGCAGGATGGAG
	Intron 1-2	3	-1	37,317,168	37,317,266			99	gtaagtgcacagccatcccatctat.....ataataatgtgttacgctttctcag
2	ENSMUSE00000172600	3	-1	37,317,108	37,317,167	0	0	60	AATTACAGGAACCTGAAACTCCCCAGGATGCTCACCTTCAAATTTTACTTGCCCAAGCAG
	Intron 2-3	3	-1	37,314,686	37,317,107			2,422	gtgagtgagtttctgtgtttaactg.....atggttaagcttattactcctctag
3	ENSMUSE00000172601	3	-1	37,314,539	37,314,685	0	0	147	GCCACAGAATTGAAAGATCTTCAGTGCCTAGAAGATGAACTTGGACCTCTGCGGCATGTT CTGGATTTGACTCAAAGCAAAAGCTTTCAATTGGAAGATGCTGAGAATTTCATCAGCAAT ATCAGAGTAACTGTTGTAAAACTAAAG
	Intron 3-4	3	-1	37,312,767	37,314,538			1,772	gtaaggtgttgctttatttgctaat.....cctacaattttatattctttttag
4	ENSMUSE00000172602	3	-1	37,312,271	37,312,766	0	-	496	GGCTCTGACAACACATTTGAGTGCCAATTCGATGATGAGTCAGCAACTGTGGTGGACTTT

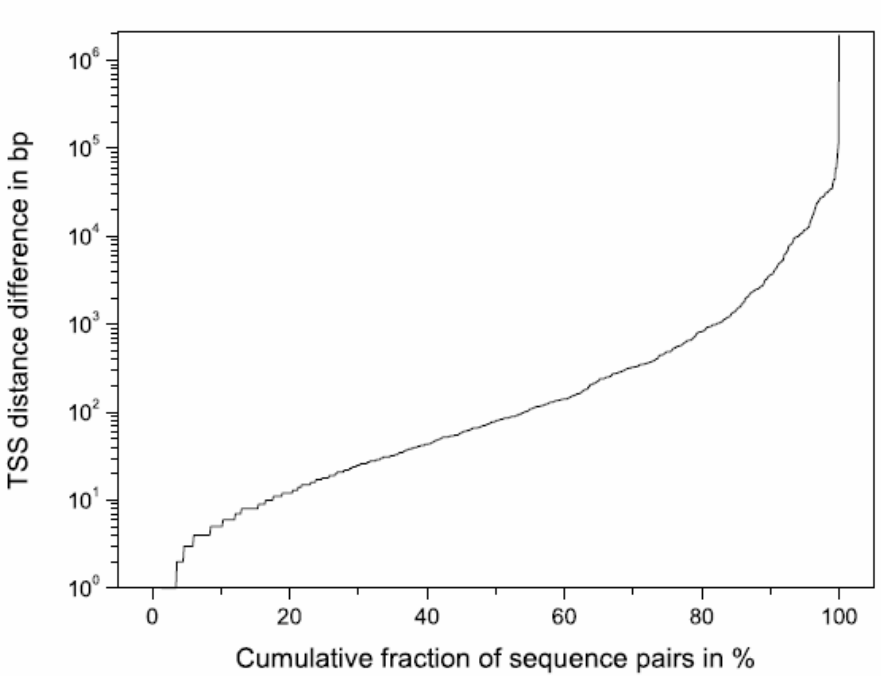
BLAST result

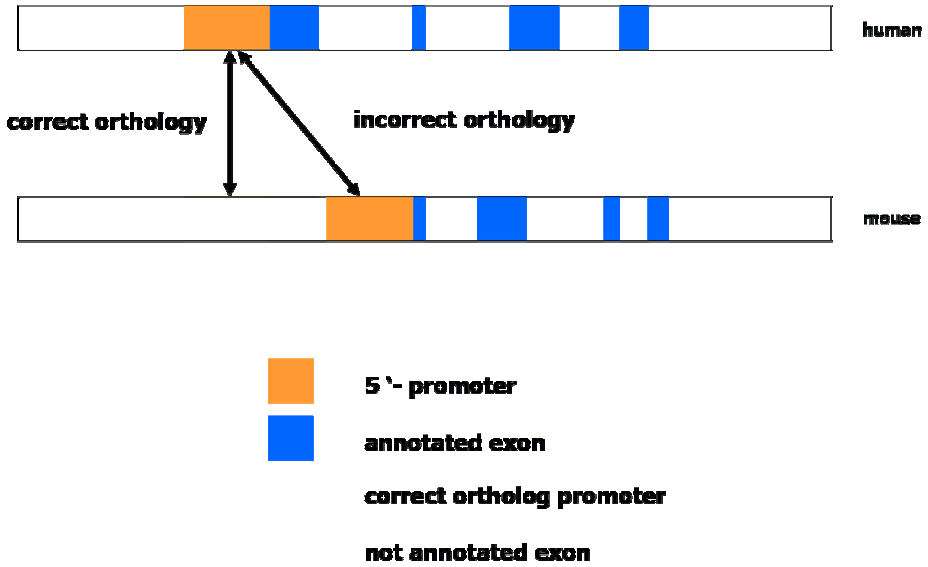
- BLAST result of the predicted human IL-2 5'-UTR against the mouse genome. The Ensembl visualization of the BLAST analysis shows that the corresponding ortholog region in the mouse genome can be reidentified with this analysis.
- The 5'-UTR region have to be extended so the promter regions have to be adapted in parallel.



Identifying true orthologs

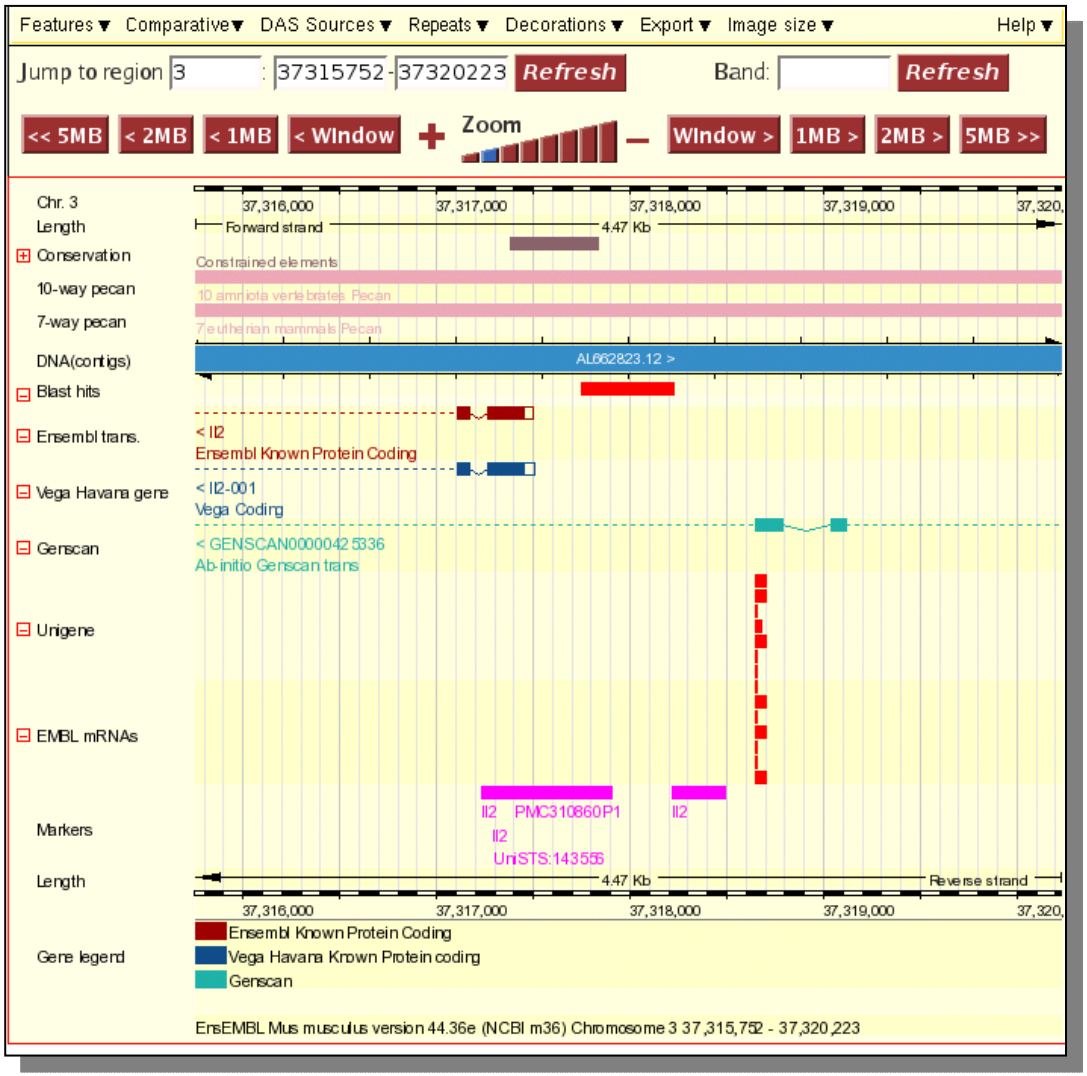
- The majority of protein-encoding genes in eukaryotic organisms starting with a 5' untranslated regions (5'-UTRs) as a first exon.
- For 775 orthologous upstream sequence pairs (human-mouse) with known TFBSs we find that ~25% of all orthologous sequence pairs differ by more than 500bp in their distance to the (annotated) TSS.



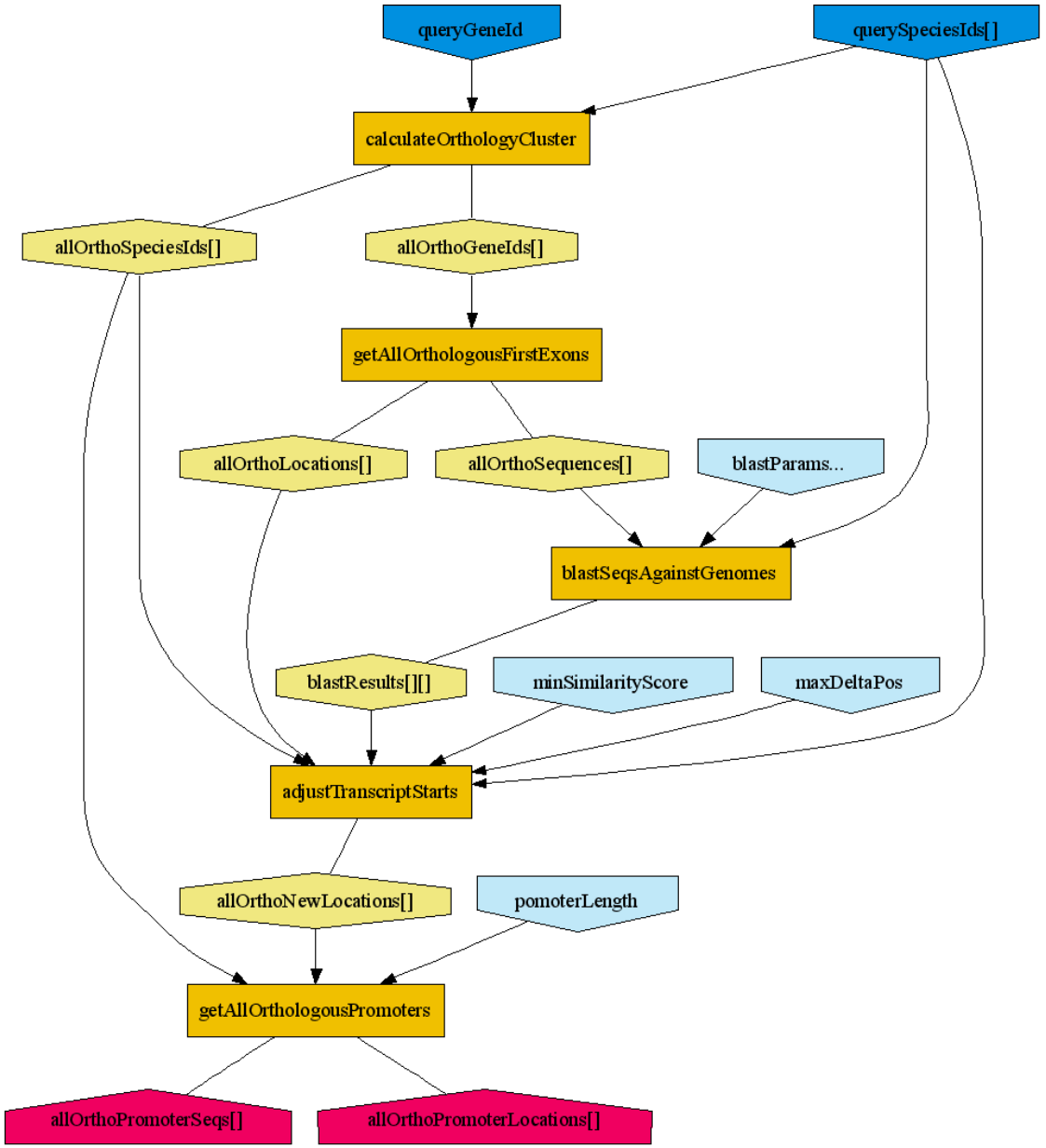


Conservation of regulatory upstream regions

- **The phylogenetic conservation of regulatory upstream regions seems to be high enough between mammalian species**
- **Blast based-reidentification within the respecitive genomes is possible**
- **Example:**
  - Blast of 500 bp human upstream promoter of IL-2 against the mouse genome
  - Alignment length: 488
  - Percent of identity: 78.07



Orthologous promoter retrieval example workflow



Requirements for workflow management systems

Requirement	Category	Mandatory?	Remarks
Conditional branching	control flow	yes	
Loop (conditional)	control flow	yes	
Loop (for)	control flow, data handling	no	Can be substituted by conditional loop + arithmetics
Loop (iteration over lists)	control flow, data handling	no	Can be substituted by for loop + by-index access
Arithmetic operators and functions	control flow, data handling	yes	
Primitive data types	data handling	yes	
Lists	data handling	yes	By-index element access, addition and removal required
Multi-dimensional lists	data structures	no	Can be substituted by one-dimensional lists + index arithmetics
Complex data types	data handling	no	Can be substituted by strings; sub-data access methods required

The presented orthologous promoter retrieval workflow defines some requirements for WMS. Roughly they can be distinguish between control flow and data handling-related requirements.

Mapping requirements to workflow management systems

- Neither of the two WMS mentioned on this slide provides all features which are required for the orthologous promoter retrieval.
- But both system are user-extensible

Requirement	Available in Taverna	Available in Bio-jETI
Conditional branching	yes	yes
Loop (conditional)	yes (implicitly)	yes
Loop (for)	yes (implicitly)	yes
Loop (iteration over lists)	yes	yes
Arithmetic operators and functions	no	no
Primitive data types	yes	yes
Lists	yes (not all required functionality available yet)	yes (not all required functionality available yet)
Multi-dimensional lists	yes (by embedding in one-dimensional-lists)	yes (by embedding in one-dimensional-lists)
Complex data types	yes (as XML, but no awareness of further semantics)	yes (as XML, but no awareness of further semantics)

Further requirements for WMS

- **Semantic process classification**
  - A classification schema (or ontology) of node types offered by a WMS is essential to identify the nodes matching a certain demand
    - Taverna: provider-oriented classification
    - Bio-jETI: definition of services taxonomies possible
- **Service transparency**
  - If the same functionality occurs multiple times in the node type list, a WMS should be able to choose the „best“ process node transparently
- **Semantic data type classification**
  - A more detailed semantic or ontology-based description of the kind of data „understood“ by the various available processing node types would be beneficial for the workflow design process (model checking)

Further requirements for WMS

- **Nested workflows**
  - Encapsulation of sub-workflow in a single, re-usable processing node. Both Taverna and Bio-jETI can collapse parts of the workflow graph into single nodes.

- **Publication support**
  - Publication of workflows to the public
    - Bio-jETI is able to export workflows as webservices
    - In Taverna no similar feature is found yet

- **Implementation of new process node types**
  - WMS must provide an easy-to-use framework for integrating user-supplied resources. Configurable database queries or command line execution services are available in Bio-jETI and Taverna.

Conclusions

- **Workflow management systems**
  - WMS like Taverna and Bio-jETI provide a considerable amount fo functionality required for systems biology tasks

- **Data-handling**
  - Requirement: List data type
    - adding, removing, indexing, check for exististancs which allows to add and remove elements, to determine wether or not a list contains element, and to access elements by their index would be a minimum requirement
  - Support for domain-specific complex data types
    - beneficial for workflow design and verification process (XML)

- **Data standards**
  - How to develop and establish domain-specific data type specifications, like XML schemas, so that they will actually get widely used within the community?

Acknowledgements

- **Thanks for your attention!!!**

- **UKG, Göttingen University (Medical school)**
  - Tilman Sauer
  - Knut Schwarzer
  - Torsten Crass
  - Edgar Wingender

- **Institute for Informatics (Göttingen University)**
  - Stephan Waack
  - Anna-Lena Lamprecht

- **Special thanks to the initiators of the part ‚From components to Processes in Bioinformatics‘**
  - Tiziana Margaria
  - Bernhard Steffen
  - Robert Giegerich