Genome Annotation without Genes

Jan Engelhardt[1,2], Toralf Kirsten[2], Peter F. Stadler[1−6], and Sonja J. Prohaska[1,2]
[1]Bioinformatics Group, Department of Computer Science,
University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany,
[2]Interdisciplinary Center for Bioinformatics, and
[3]Max-Planck-Institute for Mathematics in the Sciences,
Inselstraße 22, D-04103 Leipzig, Germany,
[4]RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology,
Perlickstraß 1, D-04103 Leipzig, Germany
[5]Institute for Theoretical Chemistry, University of Vienna,
Währingerstrasse 17, A-1090 Vienna, Austria
[6]The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM, U.S.A.

{jane,studla,sonja}@bioinf.uni-leipzig.de, tkirsten@izbi.uni-leipzig.de

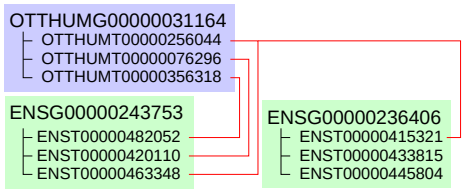
Abstract: The concept of the gene plays a fundamental role both in the interpretation and in the organization and storage of molecular biology data. Genes are treated as if they were unambiguously characterized physical entities with clearly defined measurable properties. The ubiquitous usage of genes in current bio-databases reinforces this perception. A closer look at the data models and the ongoing discussion of the gene concept itself, however, exposes the gene as an ill-defined *ad hoc* construct that is unsuitable both as interface between functional annotation and sequence level data and as organizing principle in molecular biology. It should thus be abandoned in the context of genome annotation. As a collective of associated DNA, RNA, and protein sequences, the gene should be replaced with a more a more explicit model of the expression and processing cascade, implying that functional annotation should be linked explicitly to physical objects only.

1 What is a Gene?

The notion of the **gene** plays *the* central role in present-day computational biology. A large part of the information collected over decades of research in molecular biology is organized and stored in terms of genes. Entrez Gene, NCBI's database for gene-specific information [MOPT07], makes this most obvious. Its GeneIDs serve as hubs for multiple types of information (nomenclature, gene products and their attributes, markers, phenotypes and links to citations, sequences, variation details, maps, expression, homologs, protein domains) and as a means of linking to external resources. AmiGO [CIM$^+$09] provides access to functional annotations in terms of the controlled vocabularies of three distinct gene ontologies [Gen00]. Here, records either refer to a gene or to a gene product, usually a protein, depending on the original data source. OMIM, the comprehensive compendium of human genes and genetic phenotypes, is another example, which is of fundamental importance in medical research. The fundamental role of the gene concept in organizing the

#	ENS→OTT	OTT→ENS
1	74955	75125
2	190	132
3	18	9
4	5	–
7	1	–

Frequency of multiple images in mapping between ENSEMBL and OTTER transcripts in human



MHC class I, L (pseudogene) Chr.6: 30,227,361-30,260,791 (+)

Figure 1: Inconsistencies between otter and ENSEMBL. Surprisingly, the map between transcripts contains multiple ENSEMBL genes corresponding to the same otter and *vice versa*. As a consequence, the map between the genes is also many-to-many.

overwhelming part of molecular data is further underscored by organizational structures such as the HUGO Gene Nomenclature Committee that have been established for the sole purpose of regulating naming conventions for genes.

The tight links between genes and all the different kinds of functional annotation data reflects that the currently utilized data models *implicitly* treat genes as units of functions. None of the major databases, however, gives a clear definition of the term *gene* or a theoretical framework describing the relationships of sequence data and function. Instead, expert curators bring genes into existence in case-by-case decisions, applying pragmatic procedures that regulate how GeneIDs are assigned and what "belongs to" a gene. While researchers in the field assume that they will "recognize a gene when they see one", there is, at present, no unambiguous scientific way to make this decision — an there is no general consensus, either. RefSeqGene, a part of Entrez, for example, defines "genomic sequences of well-characterized genes" by identifying one or (a few) *representative* mRNAs and associated proteins; the underlying RefSeq entries in turn are selected by curators to be *representative* transcripts and (mostly) translation products. It appears that the process involves the implicit assumption that the curation process establishes the desired link between sequence and function. ENSEMBL, on the other hand, views genes as *collections* of transcripts with overlapping coding sequence. Transcripts that belong to the same gene ID thus may differ substantially in sequence and properties of the resulting proteins.

Different data source (e.g., ENSEMBL, OTTER, TrEMBL, Entrez) typically use their own accession number nomenclature, and hence implicitly, their own gene definition. In each system, genes are characterized by associated data (e.g. DNA sequences, RNA transcripts, genomic coordinates) that are used to link accession numbers. The consistency of different data sources has to be checked empirically. Fig. 1 exhibits some inconsistencies between otter [SGIC04] and Ensembl. Even though the discrepancies are fairly rare, they imply that functional information linked gene IDs of different annotation systems cannot always be transferred.

The concept of the *gene* has come under intense scrutiny in recent years in response to the recognition that the "standard" model that views genes as beads on a genomic DNA string is inconsistent with the results of the systematic investigation of eukaryotic transcriptomes. The major issues include the following: (1) There is a very large number of well-characterized transcripts that do not code for proteins [ENC07, FAN05], and there is mounting evidence that many of them are functional. (2) Protein-coding location also produce non-coding transcripts that share coding and/or non-coding exons

Figure 2: Simplified map of the HoxAa locus of zebrafish (*Danio rerio*) genome. The three genes *HoxA9a*, *HoxA4a*, and *HoxA3a* have transcripts that overlap in exons.

with protein-coding isoforms. (3) Chimeric or mosaic transcripts produce protein coding mRNAs that combine exons sampled from what is typically thought of as separate genes [ATE$^+$06, PRD$^+$06]. (4) Different processing stages may give rise to distinct functional gene products with unrelated functions, as in the case of snoRNA-derived microRNAs or structured RNAs that also encode peptides. (5) Trans-splicing is common throughout eukaryotes, although is its rare in mammals. Trans-spliced products may arise from a non-contiguous genomic source [Gin09]. (6) Many eukaryotes, however, extensively process their DNA throughout their life cycles. Ciliates, with their separation of micronucleus and macro-nucleus are the prime example. Among vertebrates, lamprey undergoes a dramatic remodeling of its genome, resulting in the elimination of hundreds of millions of base pairs from many somatic cell lineages [SAEA09].

The rules that identify transcripts belonging to the same gene were established with splice-variants of mRNAs in mind. The focus on common exons, therefore, reflects an implicitly protein-centric point of view that considers the mature mRNAs as the most relevant stage in the life of a transcript. Many of the mRNA-like ncRNAs, however, are mere intermediates on the way to relevant RNA molecules, as in the case of microRNAs, piRNAs, and a plethora of other small RNAs [KCD$^+$07]. On the other hand, many transcripts remain unspliced and retained in the nucleus, where they have functions e.g. in the organization of nuclear structures such as MALAT-1 and MENβ. One could of course *define* such transcripts as exonic, although they have little in common with processed mRNAs. Stipulating that only the exonic parts counts, however, one has to answer in what sense ncRNAs processed from introns can be regarded as genes. In ENSEMBL, snoRNAs and intronic microRNAs are annotated as genes. The observation that protein coding loci are covered by complex superpositions of primary transcripts [ENC07] adds another level of complications. On the one hand, disjoint proteins may be produced from the same primary transcript and the same promoter, while multiple promoters produce alternative transcripts yielding the same protein product. A good example is the HoxB3a/HoxB3b locus of zebrafish [HPP$^+$06]. The current data model, finally, views the genomic DNA as a safe and stable reference (except for SNPs and copy number variations). This is true for the best-studied model organisms but not even a good approximation in many other cases.

These facts undermine the gene-centric approach in two ways. Accepting the fact that there are many functional genes that do not produce proteins, we have to abandon the protein-centric rules for assigning gene IDs and combining transcripts. For instance, we might relax the requirement to share coding exons and define a gene as a collection of transcripts that share exons. Often, this would lump together time-honored genes Fig. 2. Since the number of verified transcripts is still rapidly increasing, the connected components of exon-overlapping transcripts are also increasing, eventually covering large genomic regions or even entire chromosomes. In fact, the current ENSEMBL annotation is already making exceptions from the overlapping coding exons rule to maintain previously

annotated genes as separate entities: Mosaic transcripts that would link two previously annotated coding genes are assigned to only one of these genes despite overlapping coding exons (ATP5O-013, for instance, shares two coding exons with DONSON, [ENC07]). This is not an exotic phenomenon: Hundreds of chimeric transcription-mediated fusions of adjacent genes exist in human tissues [ATE+06, PRD+06]. The current version 57 of ENSEMBL Homo sapiens genes lists 142 transcripts groups that connect two or more ENSEMBL genes via coding exons. Combining Ensembl genes whenever exons of recognized ENSEMBL transcripts overlap would reduce the gene number by 4329 or 8.3%. On the other hand, there are 38 ENSEMBL genes in which two or more non-overlapping groups of coding transcripts are connected only by non-coding isoforms. Transsplicing causes additional problems since it violates the assumption that mRNAs are composed of a genomically co-linear sequence of exons. Transsplicing is a regular process in many species, including *C. elegans*. Various exotic exon arrangements have been observed. In the *Drosophila* mod(mdg4) gene, e.g., exons transcribed from both reading directions are combined [DRL01]. Even interchromosomal trans-splicing might not be a rare phenomenon [Gin09].

The second, and more critical, problem is the lack of functional coherence in a gene that produced multiple functional products via largely or even completely separate processing pathways. Of course, biologists have always dealt with the fact that the same protein can have different functions in different contexts – the issue at hand, however, is that our "genes" have turned out to give rise not only to closely related isoforms with arguably closely related functions, but rather to multiple products that can be of different biotypes, localize differently in the cell, take part in completely different molecular interactions, and affect unrelated biological processes. The only commonality that is left is an origin from the same genomic location.

The concept of the gene is also the subject of a debate among theoreticians in the life sciences, aiming at a reconciliation of classical notions of genes as heritable units of functions and transmission with sequence-based constructions favoured in molecular biology. Some of the dissenting opinions are expressed in a recent special issue of *Th. Biosci.* (128(3), 2009). The current state of this ongoing debate and the physical evidence outlined in the previous paragraphs inevitable lead to two conclusions: (1) There is no commonly accepted concept of the gene that is consistent with current knowledge on genome organization. (2) The association of a common function to collections of transcripts that are defined by overlap or overlapping genomic location is scientifically untenable. Consequently, "genes" are **not** an appropriate basis to organize the knowledge of molecular biology,

2 Consequences

These issues have been known for years. Why, then, are our data still organized as they are, and why should one attempt to change this now?

There are several answers to the first question. Biologists have been trained to think in terms of genes, even if the meaning of "gene" is highly context dependent and fluid in practise. In the wet-lab, the intricacies of the gene concept and the data models of biodatabases are of little practical consequence: experiments are planned and conducted in

terms of the physical entities (DNA sequences, transcripts, proteins), usually based upon the published physical evidence that underlies the functional annotation. A second set of reasons is very pragmatic: it is hard to change the current practise because it is well established, used by a large community, and has proved hugely useful. While we know about the flaws and limitations of the gene concept and its application in genome annotation, a full-fledged alternative data model does not appear to be around. All the complications discussed above, finally, are often relegated to the status of "rare and unimportant exception".

In the last couple of years, however, it has become clear beyond reasonable doubt that the "exceptions" are not rare; in all likelihood they are even more frequent than the rule [MDM09]. The success of Systems Biology, furthermore, crucially depends on the ability to utilize the available data in consistent computational model of biological reality. Systems Biology therefore needs data models that are firmly grounded in physical reality.

We argue, therefore, that data models and annotation systems in molecular need to be centered on physical entities that *can* be measured and verified unambiguously. This is pragmatically the case in all sequence data bases anyway. Their "gene" constructs can be interpreted as ancilliary data objects that bind together DNA, mRNA, and protein sequences that are in some sense related with each other. The necessary change, therefore, does not affect the primary data but the logical structure built upon them. The problematic amorphous containers called "gene" should be replaced by explicit relations that model the mutual dependencies of the sequences. These relations have two aspects: they naturally describe the molecular mechanics of the information metabolism in the form of an ontology, and they have "projections" onto the sequence level, where they implement the actual transformations of transcription, splicing, editing, or translation. One of the many advantages of such a data model, which reflects biological objects *and* their connecting processes, is the possibility to explore e.g. the consequences of alternative proposals for gene definitions [SPFK09].

In a recent pilot study, we constructed a database for the in-depth annotation of HOX gene clusters, for which detailed structural an functional information. The `adHOX` system [Ran08] uses only the physical entities as annotation items and incorporates a simple, but extensible, ontology to model the relation between different types of sequence data, allowing a modification and fine-graining of the data model by introducing, e.g., distinctions between primary transcripts, spliced mRNAs, poly-adenylated mRNAs, and mature mRNAs. Sequences are related by `part-of` (UTR, exons `part-of` mRNAs; coding region `part-of` exon), `union` (ORF = `union`[coding regions], and transformations (primary transcript = `transcription`(DNA-interval), peptide = `translation`(ORF)). The system, which appears to be perfectly workable, at present does not include any notion of a "gene" at all.

In order to ensure compatibility with the current practise of annotation it would be easy to add an additional `element-of` relation to the ontology that could be used to define collections, include those that correspond to RefSeqGenes. There do no seem to be insurmountable problems, therefore, to restructure the existing data repositories into a form that is centered on the physical objects and their relationships rather than gene-centered. Functional annotations, however, would eventually have to be re-linked to the physical objects from which they were derived. This could be an ongoing process, however, since the ontology structure could be used to ensure back-ward compatibility to the gene-centric

view by simply propagating functional annotation along `element-of` relations.

References

[ATE+06] P Akiva, A Toporik, S Edelheit, Y Peretz, A Diber, R Shemesh, A Novik, and R Sorek. Transcription-mediated gene fusion in the human genome. *Genome Res.*, 16:30–36, 2006.

[CIM+09] S Carbon, A Ireland, C J Mungall, S Shu, B Marshall, S Lewis, AmiGO Hub, and Web Presence Working Group. AmiGO: online access to ontology and annotation data. *Bioinformatics*, 25:288–289, 2009.

[DRL01] R Dorn, G Reuter, and A Loewendorf. Transgene analysis proves mRNA trans-splicing at the complex mod(mdg4) locus in *Drosophila*. *Proc. Natl. Acad. Sci. USA*, 98:9724–9729, 2001.

[ENC07] ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447:799–816, 2007.

[FAN05] FANTOM Consortium. The Transcriptional Landscape of the Mammalian Genome. *Science*, 309:1159–1563, 2005.

[Gen00] Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nat. Genet.*, 25:25–29, 2000.

[Gin09] T R Gingeras. Implications of chimaeric non-co-linear transcripts. *Nature*, 461:206–211, 2009.

[HPP+06] T Hadrys, B Punnamoottil, M Pieper, H Kikuta, G Pezeron, T S Becker, V Prince, R Baker, and S. Rinkwitz. Conserved co-regulation and promoter sharing of hoxb3a and hoxb4a in zebrafish. *Dev Biol.*, 297:26–43, 2006.

[KCD+07] P Kapranov, J Cheng, S. Dike, D Nix, R Duttagupta, A T Willingham, P F Stadler, J Hertel, J Hackermüller, I L Hofacker, I Bell, E Cheung, J Drenkow, E Dumais, S Patel, G Helt, G Madhavan, A Piccolboni, V Sementchenko, H Tammana, and T R Gingeras. RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. *Science*, 316:1484–1488, 2007.

[MDM09] T R Mercer, M E Dinger, and J S Mattick. Long non-coding RNAs: insights into functions. *Nature Rev. Genet.*, 10:155–159, 2009.

[MOPT07] D Maglott, J Ostell, K D Pruitt, and T Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, 35:D26–D31, 2007.

[PRD+06] G Parra, A Reymond, N Dabbouseh, E T Dermitzakis, R Castelo, T M Thomson, S E Antonarakis, and R Guigó. Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res.*, 16:37–44, 2006.

[Ran08] M Ranisch. adHox – A Hox cluster annotation database. Master's thesis, Univ. Leipzig, 2008.

[SAEA09] J J Smith, F Antonacci, E E Eichler, and C T Amemiya. Programmed loss of millions of base pairs from a vertebrate genome. *Proc Natl Acad Sci USA*, 106:11212–11217, 2009.

[SGIC04] S M Searle, J Gilbert, V Iyer, and M Clamp. The otter annotation system. *Genome Res.*, 14:963–970, 2004.

[SPFK09] P F Stadler, S J Prohaska, C V Forst, and D C Krakauer. Defining Genes: A Computational Framework. *Th. Biosci.*, 128:165–170, 2009.