

Biocuration: Best Practices

Monica Munoz-Torres

Elsik Computational Genomics Lab | Georgetown University
NESCent – NGS | August 18, 2011

Overview

- “ What ‘annotating’ means
- “ The naked genome
- “ Curation, defined
- “ What to look for
- “ How it’s done
- “ Annotation Tools
- “ Keep in mind

“ Annotation is a methodology to add information to a document, anchored to a specific point in the document.

Data about Data = Metadata

What 'annotating' means



Data about Data = Metadata

What 'annotating' means

“ Tags specify ranges of text in a document:
linguistics, semantics, business intelligence

“ Annotating is to connect a subject to broader concepts stored in knowledge bases or ontologies

Data about Data = Metadata

What 'annotating' means

- “ Extrinsicly: mRNA, proteins from reference genome
- “ *Ab initio*: signals & properties of the sequence
 - “ Promoters
 - “ ORFs
 - “ CpG Islands
 - “ PolyA tails
 - “ Splice sites ...]5' GT / AG 3' [...
 - “ Probabilistic methods: e.g: HMM

Gene prediction identifies elements of the genome

The naked genome:
Automated Structural
Annotation

“ Adding experimental evidence identifies domains & motifs:

“ DNA: ESTs, cDNA, RNAseq demonstrate transcription, boundaries, alt. transcripts (ESTs)

“ Proteins: provide alignments by structural similarity

Automated Structural
Annotation (Ctd’)

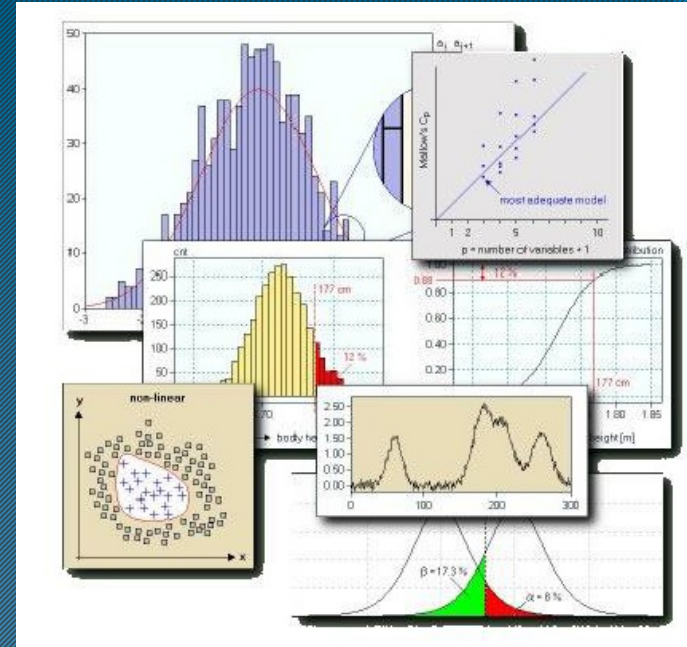
“ To find the best examples and/or eliminate (most) errors.

“ Sometimes there is no way to determine right or wrong: keep them all

Curation, defined

What is BioCuration?

The experimental research community generates data



New Research Hypothesis



Biocurators, software and database developers organize the data and make it available for users

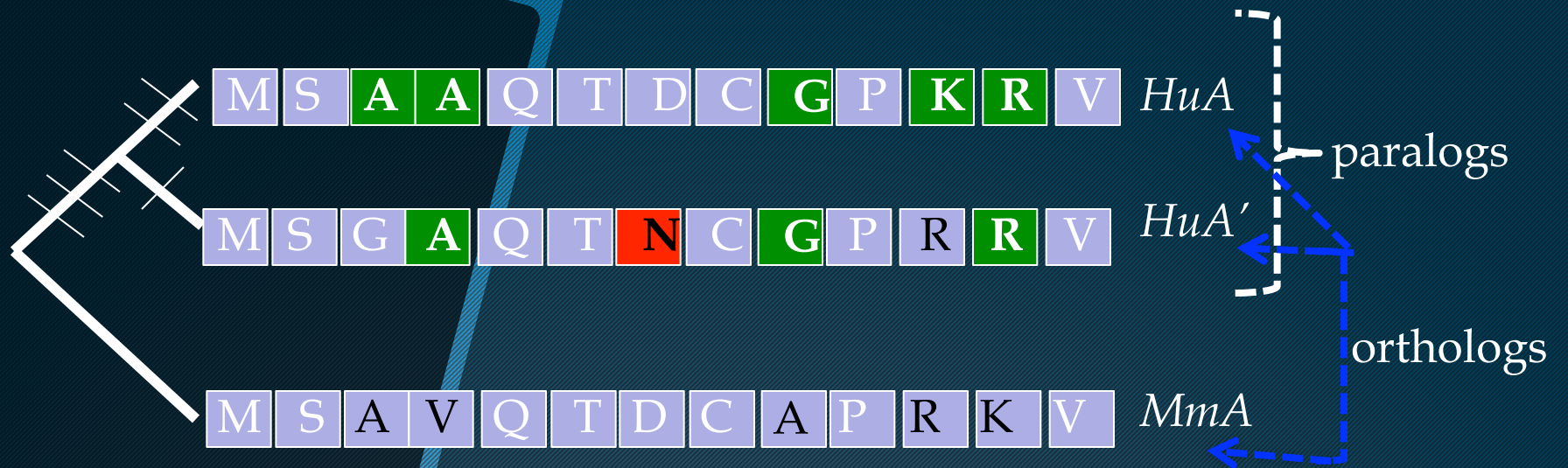
- “ Is used to evaluate all available evidence and corroborate/modify automated predictions
- “ Is done gathering supporting evidence, using quality-control metrics
- “ Is necessary because incorrect and incomplete genome annotations will poison every experiment that uses them

Manual Curation

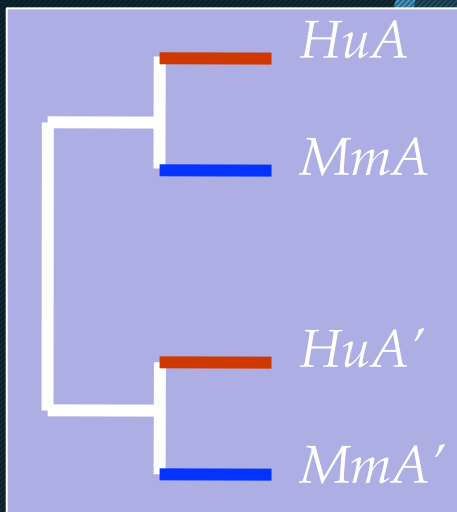
- “ Uses literature and public databases to infer gene function from experimental data
- “ Performs sequence-similarity searches within a phylogenetic framework (e.g: alignment trees):
 - “ To predict protein functional assignments
 - “ Distinguish orthologs from paralogs to classify genes as members of a family

Manual Curation

Orthologs and Paralogs

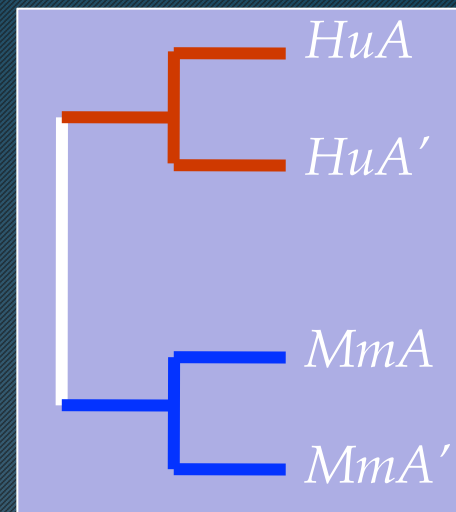


if



orthologs

if



paralogs

All methods begin with automated annotation, and then...

1. Small groups of highly trained experts
2. A few good biologists, a few good bioinformaticians camping together
3. Everyone for themselves, then submit
4. Distributed annotation, no updates

Manual Curation:
Different Methods

Dispersed Community Curation

“ Exploits biologists’ expertise:
functional annotation

“ Initially focuses on families of
interest

“ Wet lab work for additional
sequence and expression data

Manual Curation: The
Social Experiment



“ Intron/exon boundaries

“ ORFs

“ Missing alternative splicing forms

“ Frame-shift errors

“ Missing untranslated regions

“ Merged or split genes

“ Degenerate transposons annotated as protein-coding genes

“ Sequence gaps, missing 5' or 3' sequences

“ Single base errors

“ Selenocysteine (Se-Cys, no P)

“ Readthrough transcripts

The good, the bad, the ugly and
other inconvenient phenomena

What to look for

The J. Craig Venter Institute:

“ Structural and functional

“ Designed to yield rich content and high quality automated annotation

“ Individual protein annotation

“ Characterized Protein Database: manual

“ Trusted protein families TIGRFAM, Pfam

How it's done: Prokaryotic and metagenomic at JCVI

CHAR:

“ Literature curation

“ Standardized nomenclature

“ GO assignments: function, process, evidence codes, functional protein code, Enzyme Commission #, transport classification, gene symbol, synonyms

CHAR + TIGRFAMs =
accurate homology-based
functional assignments

Wellcome Trust Sanger Institute: **Havana**,
manual only

“ Transcriptional evidence

“ Controlled vocabularies: known, novel,
putative, NMD (nonsense-mediated
decay)

“ Transcripts w/o CDS: retained intron?
artifact? Putative?

“ 7 types of pseudogenes

How it's done: Eukaryotic
vertebrates at WTSI

WTSI: The **Havana Group**, manual only

“ All exons are supported (proteins, mRNA, ESTs)

“ No variants combining exons in patterns that may not occur *in vivo*

“ If & which CDS to annotate in alternative splicing? Must account for NMD, translation mechanics, conservation & domains

How it's done: Eukaryotic vertebrates at WTSI

WTSI: The Havana Group

“ Illumina transcriptomes (RNAseq) mapped to reference genome: alternative splicing at single-bp resolution

“ Exonerate / PASA alignments of ESTs, full-length cDNAs

“ tBLASTx: new gene models, exon boundaries

“ Rfam to other spp; GO; Interpro; Uniprot; orthoMCL

How it's done: Pathogens of Eukaryotes at WTSI



“ Generic Model Organism Database project (GMOD): more genome-browsing & editing tools

“ RNA sequencing technology: transcriptome evidence!

“ Yet more robust and more intuitive tools are needed to visualize, edit, analyze and annotate genes, gene products, features and attributes.

An old story's newest toys

- “ JCVI’s Manatee
- “ WTSI’s Otterlace & Zmap
- “ WTSI’s Artemis
- “ GMOD’s Apollo

Annotation Tools

- “ free, open-source, web-based
- “ secure remote login, or local
- “ GUI
- “ multiple annotations simultaneously
- “ write-backs
- “ Multi-Genomic Annotation Tool (MGAT):
evidence, synteny, propagation of information
across protein clusters of different species

Manatee

A

GENE CURATION INFORMATION


ORF04813 (SO2740)
 ▶ View BER Searches
 asmbi_id: 7974
 ▶ Reload Page

end5/end3: 2856763 / 2855711
 gene length: 1053
 protein length: 350
 molecular wt: 38790.13


database:
 feat_name / locus:
 New Gene

Select Display
 Select Function
 Refresh Searches

GENE IDENTIFICATION

gene name:
 gene_sym:
 EC number(s): 
 private comment:
 public comment:

EC GO suggestions:
 ▶ GO:0004076 biotin synthase activity (F)

submit | history | 


nt_comment

B

EVIDENCE PICTURE

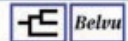


sec structure: Coil(-), Strand(blue), Helix(yellow)
 S02740
 TIGR00433: biotin synthase
 PF06968: Biotin and Thiamin Synthesis associated domain
 PF04055: radical SAM domain protein
 COG0502 (p-value: none)
 Characterized match: SP:P12996

submit | 

C

BER SKIM

 View BER Searches search date: Wed Oct 23 12:59:20 2002 Refresh Searches

accession	% sim	length	description	p-value
OMNI-SO2740	100.0	349	biotin synthase (Shewanella oneidensis MR-1)	1.5e-176
			ynthase (EC 2.8.1.6) (Biotin synthetase), [Serratia	2.5e-119
			ynthase (EC 2.8.1.6) (Biotin synthetase), [Escherich	7.2e-120
			biotin synthetase (Escherichia coli)	1.5e-119
			ynthase BioB (uncultured bacterium pCosHE2)	1.5e-119
			etase (Escherichia coli O157:H7 VT2-Sakai) [CGP13	5.1e-119
			se (Yersinia pestis CO92) [OMNIINTL02YP2986 biot	8.3e-119
			-like protein (uncultured bacterium pCosFS1)	9.5e-118
			nthesis, sulfur insertion? (Escherichia coli O157:H	2.2e-118
			ynthase (EC 2.8.1.6) (Biotin synthetase), (Erwinia h	3.6e-118
			ynthase (EC 2.8.1.6) (Biotin synthetase), [Salmonell	5.1e-119
			ase (Vibrio cholerae El Tor N16961) [CGP9655583]g	5.1e-119
			etase (Salmonella enterica serovar Typhi CT18) [CG	1.1e-118
			se (Pseudomonas aeruginosa PAO1) [CGP9946364]gbl	7.7e-116
			ynthase BioB (uncultured bacterium pCosAS1)	9.1e-113
			ase (Xanthomonas campestris pv. campestris ATCC3	2.8e-111
			ase (Xanthomonas axonopodis pv. citri 306) [CGP21	6.6e-110
			se (Buchnera aphidicola Sg) [CGP21623185]gbl/AAM6	1.4e-109
			ase (Xylella fastidiosa 9a5c) [CGP9104834]gbl/AAF8	8.4e-110
			TIN SYNTHASE PROTEIN (Ralstonia solanacearum GMI	4.7e-109
			ynthase (EC 2.8.1.6) (Biotin synthetase), [Buchnera	1.1e-107
			otin synthase (Acinetobacter calcoaceticus)	1.6e-106
			ase (Caulobacter crescentus CB15) [CGP13425251]gb	3.0e-105
			LASE (Brucella melitensis 16M) [CGP17984969]gbl/AAL	6.3e-105

Manatee

Otterlace & Zmap:

“ BACs BLAST vs. all available DBs; send results to MySQL, visualize using Zmap. Bixrem & Dotter build alignments, Lace for transcript editing.

Artemis:

“ Free; all OSs; adapted to Chado.

“ Allows simultaneous view of multiple sequence alignments in the context of a genome

“ Artemis Comparison Tool = whole genomes

WSTI's Tools

Artemis File Entries Select View Goto Edit Create Run Graph Display

Artemis Entry Edit: MAL1.embl

Entry: MAL1.embl

10 selected bases on forward strand: 299715..299724

GC Content (%) Window size: 120

User algorithm from MAL1.3.timepoints.plot Window size: 16

>> <<

PICEN1 PFA0355w:exon:1

293600 294400 295200 296000 296800 297600 298400 299200

PFA0350c

<<

```

L I E F K I T T L Y C F L D F I M I R M * I T Q R K # S Y Q K I K K K
L + N L K # L H Y I A S S I L S * # G C E L L K G S N H I K K # K K
S Y R I # N N Y I I L L P R F Y H D K D V N Y S K E V I I S K N K K K
C T T A T A G A A T T A A A A A A C T A C A T T A T T G C T T C C T C G A T T T T A T C A T G A T A A G G A T G T G A A T T A C T C A A G G A A G T A A T C A T A T C A A A A A A A A A A A A A A
297100 297120 297140 297160 297180 2972
G A A T A T C T T A A A T T T A T T G A T G T A A T A T A A C G A A G G A G C T A A A A T A G T A C T A T T C C T A C A C T T A A T G A G T T C C T T C A T A G T A T A G T T T T T A T T T T T T T
K Y F K F Y S C # I A E E I K D H Y P H S N S L P L L * I L F Y F F F
# L I # F L + M I N S G R N # * S L S T F # E F S T I M D F F L F F
R I S N L I V V N Y Q K R S K I M I L I H I V * L F Y D Y * F I F F F
  
```

<<

PFA0345w:mRNA	293860	295186	centrin-1
PFA0350c	295518	297149	c conserved Plasmodium protein, unknown function
PFA0355w	297580	299611	carbon catabolite repressor protein 4, putative

Artemis WSTI's Tools

Berkeley- Drosophila & WTSI | GMOD

- “ Allows easy generation and update of gene models and exon boundaries based on overlapping (stacked up) evidence sets
- “ Selecting one gene model highlights all overlapping evidence
- “ Exon-detail editor: click and drag
- “ ‘Sequence Aligner’ & ‘Align selected features’: color-coded alignments; detection of pseudogenes, structural edits; highlights differences between expressed and genomic sequences.

Apollo

“ some re-annotate and curate an already annotated genome

“ some annotate many genomes *de novo*

“ some **also** use literature

“ some focus on evidence-based functional annotation

Keep in mind

Different styles...

- “ Choose an optimal computational gene prediction tool
- “ Improve existing genome annotations & choose a strategy for regular genome updates
- “ Keep track of changes in gene updates and supporting evidence;
- “ Stay in the know of changes in our understanding of biotechnology and gene biochemistry.

Madupu et al, DATABASE. 2010 doi:
10.1093/database/baq001

...benefit from
common starting
points

“ Prioritize genes for curation:
most likely to be incorrect?
Specially interesting? by
specific type?

“ Document all changes:
priorities, algorithms,
standards, which datasets for
annotation you choose, and
SOPs

Madupu et al, DATABASE. 2010 doi:
10.1093/database/baq001

...benefit from
common starting
points

“ The community-based curation model is the most accepted for large genome projects.

Join the
conversation!

“ Thank you.

Questions?