A human-aided genome annotation pipeline

Francis Ouellette
Director, UBC Bioinformatics Centre
francis@bioinformatics.ubc.ca
Outline

• VanBUG (Stef)
• UBiC: MyPlan
  • MyPipeline
    – What we need
    – What we have
    – What is wrong with what we have
    – How we will get there
• A short comment on Open Source
Bioinformatics is about understanding how life works. It is an hypothesis driven science.
In bioinformatics, we use software tools and biological databases to ask questions.
At the UBC Bioinformatics Centre (UBiC) we bring together scientists that share the vision of making advances in computational biology, also working with bench scientists to validate the hypotheses we are generating.
UBiC: the vision

- Basic Research
- Large Scale Bioinformatics
- Support & Training
MyPlan

• Building a Bioinformatics Centre at UBC
  – BC is the most fertile ground in Canada for doing this.
  – Leverage this against the large scale genomics and proteomics efforts in Vancouver, and worldwide.

• Build a BC focal point where bioinformatics, genomics and proteomics can be integrated in one Centre.
  – Be part of the life-sciences community at UBC and work with them to advance science.

• Serve a community of about 2,000 scientists in multiple faculty & departments.
  – Do this without diminishing the kind of service that has been offered to CMMT scientist in the last 4 years.
Structure

- Director
- Associate Director
- 6 adjunct faculty
- 4 more to be recruited
- Another recruitment already in progress

- Director of Operation and Strategy
- Director of Finance
- Chief Soft. Dev.
- Chief Bioinformatics
- Chief Systems
- Chief Training and Support
- Chief Web Development
The UBiC (adjunct) Faculty

- Dave Baillie
- Jenny Bryan
- Anne Condon
- Holger Hoos
- Steve Jones
- Michael Murphy
- Francis Ouellette
- Wyeth Wasserman
- “David Wishart (UC)”
- TBD_1
- TBD_2
- TBD_3
- TBD_4
Why UBiC is special

• The people
  – Now 8 labs with some 50 people, and will grow to more than 200 in a very short time frame.

• The environment
  – CMMT/CGDN
  – GSC/BCCA
  – Joint UBC/SFU bioinformatics training program.
  – Biotechnology Laboratory
  – Beta Lab (Computer Science @ UBC)
  – SFU (Computer Science and MBB)
Ouellette Lab projects

- GeneComber: an *Ab initio* gene finding algorithm.
- IDB: the Integral DataBase system
- MyPipeline: Human-aided genome annotation pipeline
- GeMS: Genomic Mutational Signature Sequences.
- Core facility: training and support
Human-aided annotation pipeline

What we (life-scientists) need:
- An annotated (human | sea urchin | poplar | E. coli) genome that represents our best understanding of the state of knowledge for that genome.
- Current and up-to-date (at least to the day)
- Good Graphical User Interface (GUI)
- Good documentation

What developers and bioinformaticians need:
- Full access to public data and open source code
- Great GUI
- All files and formats available by anonymous FTP
- API: application programming Interface
- Documentation
When we annotate: where do we stop?

- Where?
- What?
- How?

Stein L, (2001)
Nature Review Genetics
2:493-503
Human-aided annotation pipeline

- What we have: EBI version: Ensembl

  Showing “known” (from RefSeq) and “novel” genes (from near full-length cDNA)
Human-aided annotation pipeline

• What we have: (NCBI version)

  – Many tracks and configurations possible
Problem with these Platforms:

• Conservative & not flexible
  – Current version of Ensembl: 22,980 genes shown. We know this number to be in the range of 40-60,000.
  – Ensembl is fully automated, and this does not allow user-driven input.

• Does not deal well with alternative splicing of mRNA.
  – Estimates that as much as 50% of the Human genome is alternatively spliced – less than 10% in Ensembl and NCBI’s Map viewer.

• Non-interactive, unless you are DDBJ/EMBL/GenBank
  – No published way to get your data in these systems. Databases have a hard time with what they call “3rd party annotations” or TPA (and so they should!).
What we need:

- An annotation system that allows higher throughput input into a local database so that records can now hold the generated analysis results.
- This needs to be flexible, fast and adaptable to new analysis tools and growing databases.
- Should cater to biologists, and when possible take advantage of the bio-open source community we are part of.
- This should be scalable, to be used by labs of small size (one or two people), or larger groups (10-100 people).
MyGene

All clones

All SNPs

MyGene

All mRNAs

All proteins

All structures

- All protein modifications
- Ontologies
- Interactions (complexes, pathways, networks)
- Expression (where and when, and how much)
- Evolution
Public Data
- GenBank
- RefSeq
- SwissProt
- MMDB
- BIND
- PubMed
- dbSNP

Process through suite of tools

Apollo: Annotation Tool

Validation

AnnotDB
APOLLO
Gene Annotation Tool

Michele Clamp, Suzanna Lewis,
John Richter, Steve Searle,
and nearly everyone else

Copyright BDGP
2001
Suite of Tools

• BLAST
  – Protein
  – RNA (cDNA and EST)
  – Genomic (near and far)

• Gene Finding:
  – GenScan
  – HMMGene
  – Wise2 (pseudogene)
  – GeneComber
“Parts List”

• Human genome encodes 30-60,000 genes.
• Number is even more speculative if you consider alternative splicing.
• If we are to extract knowledge from all genomes, we need to exhaustively and accurately ascertain all of the parts if we are to figure out what the underlying mechanisms of life are.
• For the identification of drug target, it is clear that having a comprehensive list is key to ensure that all relevant programs are covered.
GeneComber

- A new algorithm for the identification of likely gene products from any genome project. (Rogic et al, 2002 Bioinformatics 18(8):1034-1045)
- Probabilistic approach which takes advantage of the best from GenScan and HMMgene.
- We are in the process of making this resource available to the community.
  - Stand-alone tool
  - Testing whole genome processing
Building a tool

- Biological problem
- Development of algorithm
- Planning/modeling
- Prototype
- Productotype
- Re-engineering
- Production
- Testing
- Deployment
- Fine-tuning
- Support and documentation
Welcome to GeneComber (v0.1)

GeneComber is a graphical interface to the output of several gene prediction algorithms. For a full description of the algorithms employed please consult the online documentation.

The following are the options available to web users:

**Submit Sequences**

Submit a sequence using either a FASTA file, GenBank Accession number, or output from the GenScan and HMMGene programs.

**Retrieve Results**

Retrieve the graphical output for a sequence that you have already submitted, using the job id that you were provided with. Sequences are stored on the GeneComber server for 7 days after submission.

**Display Submissions**

Retrieve a list of your submitted sequences and their corresponding job ids.
Public Data
- GenBank
- RefSeq
- SwissProt
- MMDB
- BIND
- PubMed
- dbSNP

Process through suite of tools

Validation

Apollo: Annotation Tool

AnnotDB
Open Source

- Essential for us to exist, provides the code we use and adapt, and do the science we want to do. Millions of lines of code exist, here are some example:
  - (BLAST)
  - (NCBI toolkit)
  - Apollo
  - Perl and PHP
  - Bio-*
  - BIND software
  - GeneComber
Open Source

• In spirit, it means that you share and release source code.
• Open source takes advantage of community-based software development.
• We need to support this community, and my lab is actively doing so.
• I encourage all software developers to do so as well, academic and industry alike.
Acknowledgements:

• CMMT
  – Michael Hayden and all other faculty, PDFs and students.

• System group and Web Development:
  – Miroslav Hatas
  – Jonathan Falkowski
  – Scott McMillan

• Administration:
  – Dianne Moore

• Operations and Strategy:
  – Julie Stitt

• Bioinformatics and Software dev:
  – Stefanie Butland
  – Graeme Campbell
  – Patrick Franchini
  – David He
  – Graham McVickers
  – Jessica Sawkins
  – Sohrab Shah
  – Grace Zheng