

COMPUTATIONAL TOOLS FOR THE ANNOTATION OF NASONIA GENOME

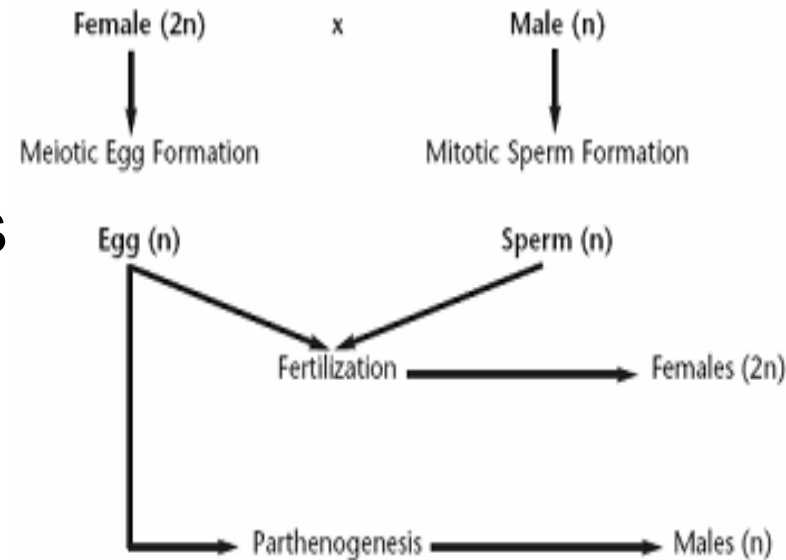
Prashanthi Selvanarayanan
Computational Biosciences
November 30, 2007

Parasitic Hymenoptera

- Model organisms for complex genetics traits
 - interesting and diverse biology
- Contains insects beneficial to humans than in any other insect group.
- Major regulators of arthropod populations in nature

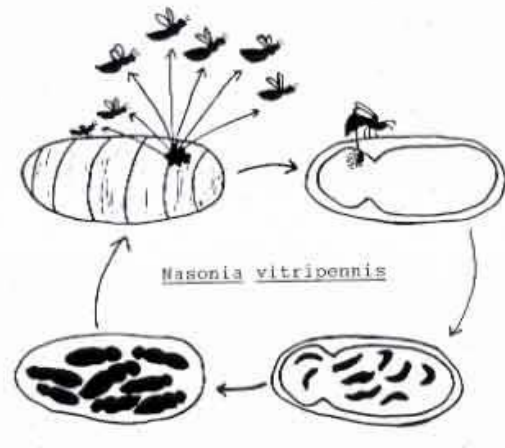
Nasonia

- Small parasitoid wasp
- Model for parasitoid genetics
- Haplodiploid reproduction
- Short generation time



Source: <http://resources.wardsci.com/tag/nasonia/>

Different Species in Nasonia



Three species:

- *N. vitripennis*,
- *N. giraulti*,
- *N. longicornis*

Source:

<http://www.biologycorner.com/worksheets/nasonia.html>,

<http://www.rochester.edu/College/BIO/labs/WerrenLab/nasonia/>

Genome Annotation

- Gene product names
- Functional characteristics of gene products

Elements Of Annotation:

- Homology searches
- Gene finding
- Functional assignment

Automated genome annotation

- Automatic annotation techniques support and complement manual curation process
- Computers do a fair job at preliminary annotation
- Requires minimal manual intervention

A Genome Annotation Pipeline- ANAP

- Goal :
To create an automated system which can deliver highly accurate and reliable genome annotations
- Essence:
Integration of suites of bioinformatics software tools
- Approach:
Combined approach of similarity search methods and gene predictors

Programming Environment

- Software was developed for this project using: Perl, Shell Script, CGI, HTML, MySQL
- The external programs utilized were: formatdb; blastall; snap; gbrowse
- The system was implemented on Mac OS

BLAST

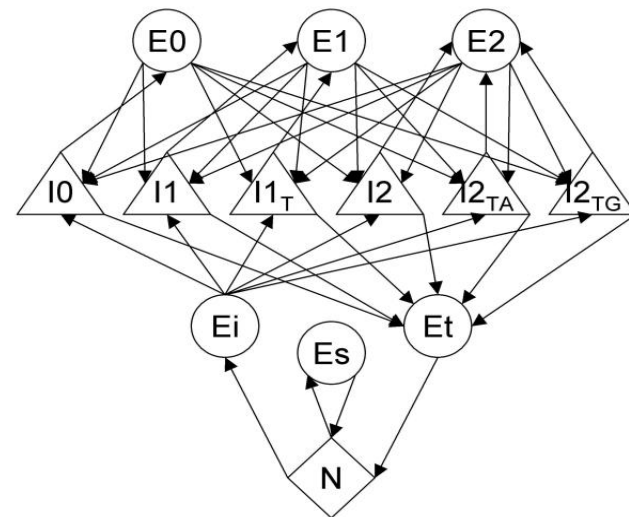
- The initial stage of computation involves homology searches using BLAST
- BLAST is a powerful tool for locating protein and cDNA sequences in the genome
- Homology results can be extended to yield an accurate gene structure
- Results filtered using e-values

Gene Prediction in new genomes

- Homology- based
 - target sequence compared against a set of library sequences from different genomes.
 - cannot not identify genes not present in the library
- *Ab Initio* gene finding
 - searches for certain signals of protein coding genes
 - complex for Eukaryotes

Semi Hidden-markov-model Nucleic Acid Parser

- A high-performance *ab initio* gene finder
- Models protein coding sequence using specialized Hidden Markov Model
- Allows changing the state diagram to describe a variety of genomic features



Source: Korf, I. "Gene Finding in Novel Genomes."

Parameter Estimation

- The solution for training a gene finder in the case of a novel genome is bootstrap Parameter Estimation
- Even the worst bootstrapped parameters are reasonably accurate for novel genomes

Source: Korf, I. "Gene Finding in Novel Genomes."

Durbin, Richard, et al. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids

SNAP for ANAP

- Trained using annotated gene sequences from *Apis Mellifera*
- SNAP is then retrained for *Nasonia* using the parameter estimation technique
- SNAP is then used to predict the genes which are in a file and used in ANAP

Gnomon Gene Prediction

- HMM-based gene-prediction used at NCBI
- Finds the maximal self-consistent set of corresponding transcript and protein alignment
- Genes predicted by Gnomon are stored in a file which is used in ANAP

Source: <http://www.ncbi.nlm.nih.gov.ezproxy1.lib.asu.edu/genome/guide/gnomon.html>

Gene Finding in ANAP

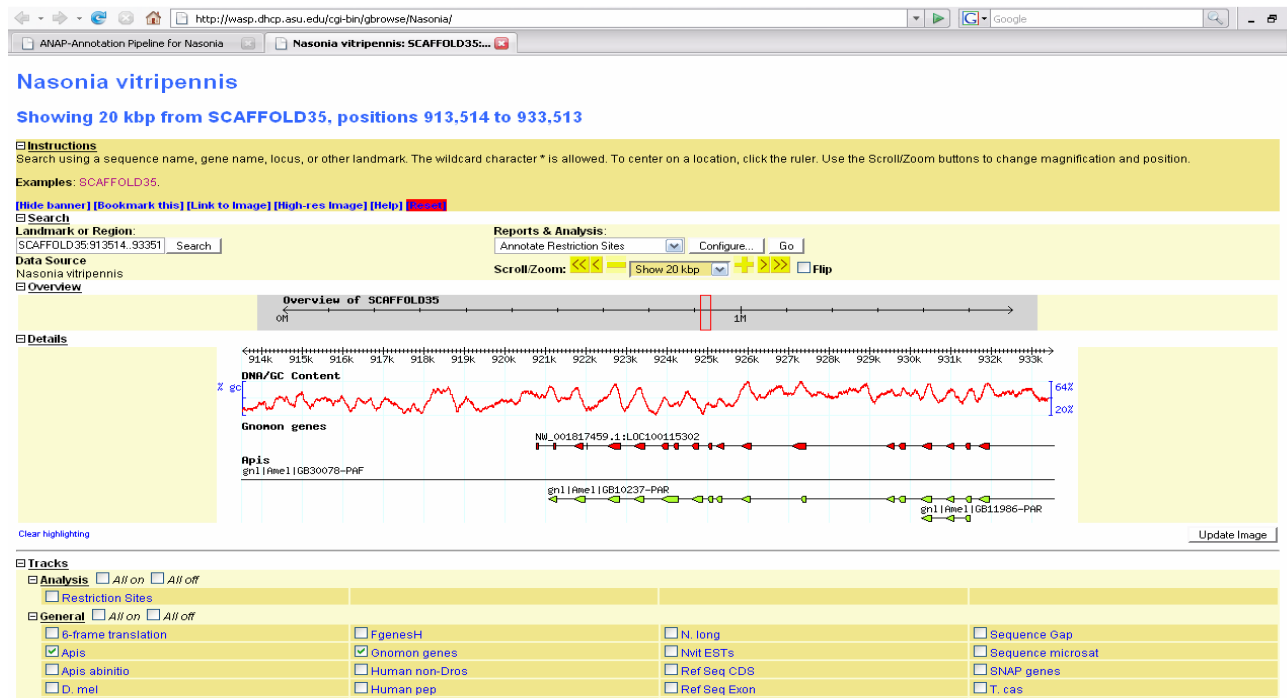
- Predicted genes are compared against the sequences resulting from homology matching
- Comparison yields two sets of outputs
 - set of Nasonia sequences with significant sequence matches with the predicted genes
 - set of Nasonia sequences with no matches with the predicted genes

Functional Annotation

- Comparative Genomics:
Annotations are based on the concept of conservation of genes across genomes of related species
- Putative gene function is based on BLASTX hits from GenBank's translated nr database
- Results are filtered using e-values

GBrowse

- A web-based display to display an arbitrary set of features on a nucleotide/protein sequence



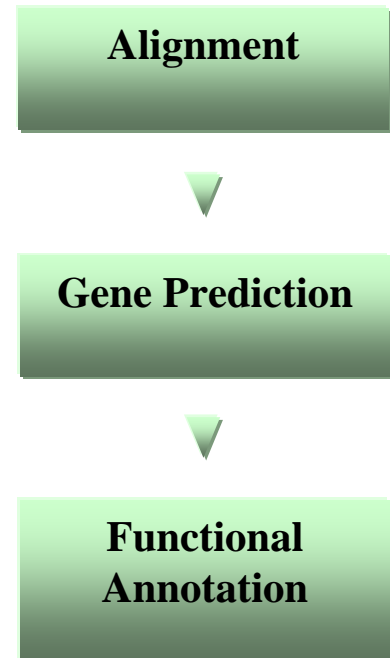
ANAP

- A web-based system
- Provides automated genome annotation
- Visualization of annotated genes

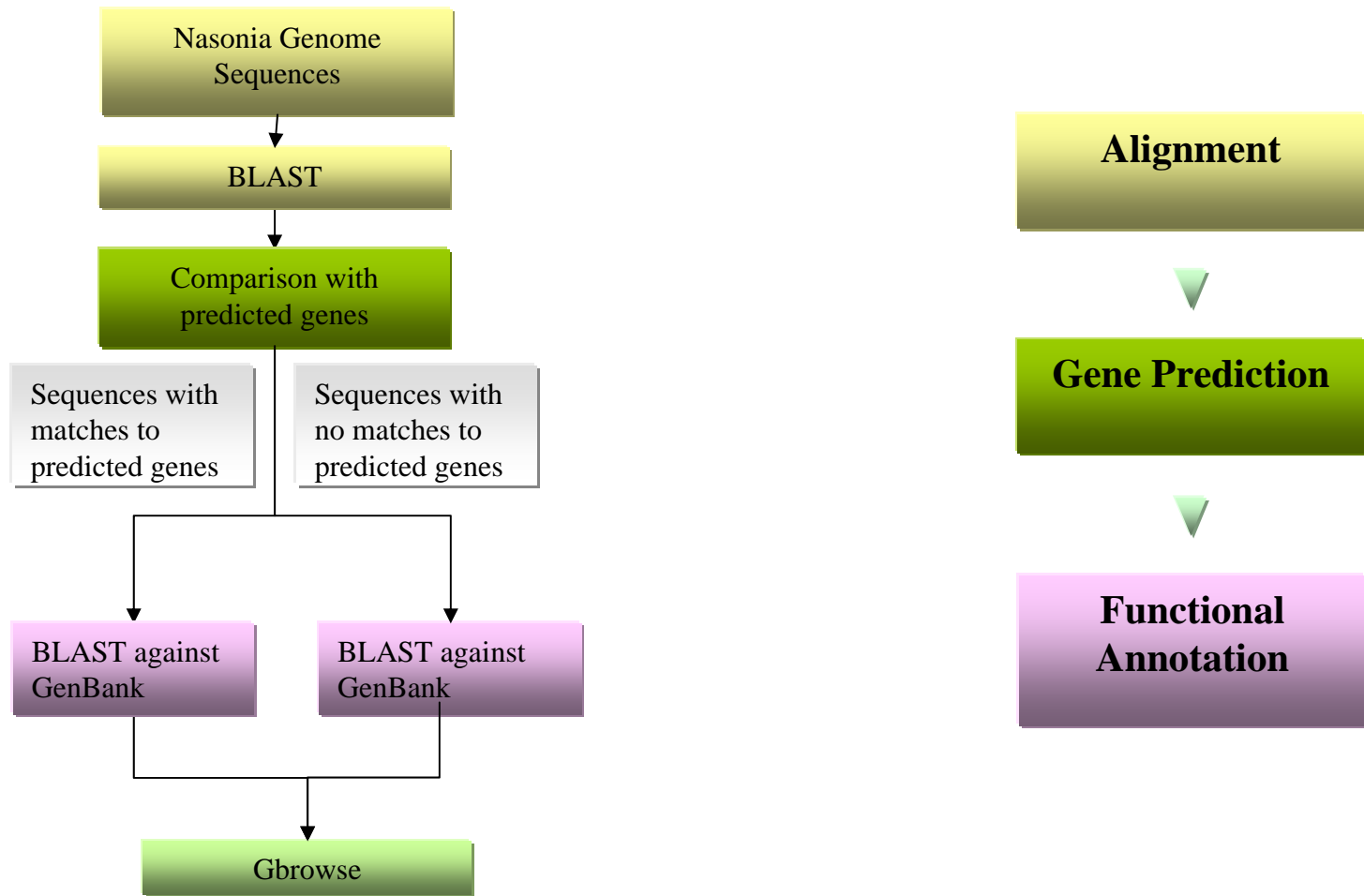
Design of ANAP

ANAP involves three consecutive steps:

- Sequence pre-processing by identification of homology matches using an alignment tool
- Gene predictions based on an ab initio gene finder
- Functional annotation based on comparative genomics



Implementation of ANAP



ANAP's Web Interface

- Allows user to input e-values for BLAST
- Allows the user to download the results from the server
- Sends email when analysis is complete
- Allows user to view the output of the annotation in a genome browser for further manual curation
- Can be accessed at <http://wasp.dhccp.asu.edu/anap/>

A Look at ANAP

ANAP- Annotation Pipeline for Nasonia Genome

Please enter a name for the job:

Please enter your Email ID:

Please upload the input file in Fasta format:

Default values have been set for all the options. Please modify the settings below to change the default values.

Enter e-value for Blast:

Select database against which the input genes should be compared:

Genbank Genes SNAP Genes

Enter e-value for Blast against Genbank:

To view your files in GBrowse: [Click here](#)



Screenshots



Copyright John Werren

ANAP- Annotation Pipeline for Nasonia Genome

Please upload the input file in Fasta format:

Select Input File Format:

Fasta GFF

Please enter the name of the sequence:

Please enter name of aggregator used in GFF file:

Specify the color for the sequence:

Please enter the Feature for the sequence:

Please enter the Key for the sequence:

Back to: [ANAP](#)

Importance of Nasonia

- Play role of biological control agents by keeping populations of insects, pests in check
- Controls invertebrate vectors of disease, like house fly benefiting human health.
- Provide additional orthologs to the human genome
- Help effective biological control of pest.

Source: Werren, J. H., et al. Proposal to Sequence the Nasonia Genome

Conclusions

- The annotation system with an easy to use interface allows users to interact with the system
- Annotation of the *Nasonia* genome will open the door to better understanding of important biological processes
- Lead to methods for control of agricultural pests and disease vectors and thus benefit human health

Future Directions

- Generic tool to support all novel genomes
 - BLAST database
 - Gene prediction

- Improve the gene predicting capability
 - Updating the predicted gene output
 - Using a more customized gene predicting software

Acknowledgements

- Dr. Jürgen Gadau and Dr. Stephen Pratt for proposing the project and providing guidance throughout.
- Dr. Zdzislaw Jackiewicz for kindly consenting to be on my committee.
- Dr. Oliver Neihuis and Josh Gibson for their support and patience.
- Dr. Rosemary Renaut for her guidance and support throughout the program.