IsgGui: A New tool for Sequence simulation

presented by
Catherine Anderson

September 2010

anderson@cse.unl.edu
Outline

- Introduction
- Sequence Simulation
- Indel-seq-gen
- IsgGui
- Discussion
- Future work
Evolution: The change in the inherited traits (as determined by changes in DNA) of a population of organisms through successive generations.

Studies investigate:

- Biodiversity: species vs. species and within species
- Pathogenicity: one species beneficial and another toxic
- Viruses: Development of vaccines
- Congenital illness: possible gene therapy
- Alternate biochemical pathways
- Drug reaction
Evolutionary hypothesis:

- From sequenced DNA
  - Genes identified (predicted or from lab work)
  - Transcripts predicted or isolated in lab
  - Protein sequence decoded and archived

- Multi-sequence alignment
  - Search in databases for similar sequences (Blast)
  - Sequences arranged into multi-sequence-alignment (MSA)
    - Manual alignment by human curator
    - Manually adjusted automated alignments
    - Fully automated alignment

- Phylogeny generation
  - Evolutionary tree predicted from MSA
  - Techniques based on maximum parsimony, maximum likelihood or other computation method.
Multiple sequence alignments (MSA)-1

3_5 exonuc (PF01612): sequence segment vs. MSA

IsgGui: A New tool for Sequence simulation presented by Catherine Anderson

Multiple sequence alignments (MSA)-1

3_5 exonuc (PF01612): sequence segment vs. MSA

IsgGui: A New tool for Sequence simulation presented by Catherine Anderson

Multiple sequence alignments (MSA)-1

3_5 exonuc (PF01612): sequence segment vs. MSA

IsgGui: A New tool for Sequence simulation presented by Catherine Anderson

Multiple sequence alignments (MSA)-1

3_5 exonuc (PF01612): sequence segment vs. MSA

IsgGui: A New tool for Sequence simulation presented by Catherine Anderson

Multiple sequence alignments (MSA)-1

3_5 exonuc (PF01612): sequence segment vs. MSA

IsgGui: A New tool for Sequence simulation presented by Catherine Anderson
Three methods

- **ClustalW**
  - Does alignment between of all pairs
  - Generates distance matrix
  - Using nearest Neighbor, builds guide tree
  - Performs alignment using guide tree

- **Mafft**
  - Converts each amino acid to a vector of volume and polarity
  - Uses Fast Fourier Transform to calculate correlation between two amino acids
  - Detects areas of similarity by peaks in FFT
  - Aligns homologous areas

- **Muscle**
  - Devises a distance measure using ‘k-mers’, based on compressed alphabets,
  - Uses a “log-expectancy” function to align profiles.
Multiple sequence alignments (MSA)-3

Pixel plots of 4 alternative alignments

- File: aaa3 fasta.ma, Coordinates: 690 to 739
- File: aaa3 fasta_muscle.ma, Coordinates: 692 to 741
- File: aaa3 fasta_mafft.ma, Coordinates: 692 to 741
- File: aaa3 clustalw2.ma, Coordinates: 695 to 744
Weakness in flow

- For extant species, exact phylogenies are not known.
- Manually generated alignments or phylogenies are best guess only.
- Evolution invoked in laboratories with mutagens provide further insight but are limited in scope [D.Hillis, 1992]
- Convergent evolution disallows phenotype guide trees.
- Automated alignments methods agree in conserved segments but vary in area of divergence.
Using Sequence Simulation

- Can generate a set of homologous sequences (all)
- Can generate the “true” MSA (most)
- Can generate the “true” phylogeny (some)
- Can generate a set of heterogeneous protein sequences (few)
- Can mix pseudogene, non-coding (intron) and coding (exon) areas in one sequence (one)

Using simulation results:

- With the “true” MSA can evaluate alignment programs and the effects of “disagreement” on the results phylogeny
- With the exact or “true” phylogenies, can evaluate phylogenetic programs
- Have a family of heterogeneous proteins
**Parameters for simulation:**

- Root sequence (DNA or protein)
- Guide tree
- Model of evolution
  - Substitution matrix (PAM, JTT, ... etc)
  - Distributions for substitution events
  - Distribution for indel events (insertion and deletion)
  - Functional constraints for proteins (clade and motif parameters)

**Results from simulation:**

- Family of taxon sequences including ancestors
- “True” multiple sequence alignment
- “True” phylogeny of sequences including mapping of indel events
- Record of indel events
Overview of iSGv2:

- Developed by Cory Strope, 2010
- Modified iSGv1 to allow generating realistic protein families.
- Can parameterize and simulate heterogeneous domains.
- Can generate divergent protein sequences without destroying functional properties.
- Combines all features offered in other simulation programs.
- Offers improved indel event generation.
Description of iSGv2:

- A command-line program
- Simulation objective can require many lengthy parameters

```
indel-seq-gen -m JTT -j des -o f -e lipocalin_out -d 011110 -s 70 -n 5 -w
  -b 1.5 -a 1.4 -g 16 -i 0.3 -z 6543 -f
  0.045,0.05,0.05,0.05,0.06,0.05,0.04,0.05,0.05,0.03,0.05,
  0.07,0.05,0.05,0.05,0.015,0.025,0.05 < lipocalin.tree
```

- Advanced parameters have dependencies
- Can be time consuming to set up and check parameters
Description of iSGv2:

- **Support file needed:**
  - substitution rates (protein or DNA)
  - indel occurrence rates
  - indel length distribution rates
  - lineage file (clades and motifs)

- **Guide tree file:**

```
[:lipocalin_ma(1:56,6,c)]"motif_1" {5,0.1,idLD}(((taxon1:0.32,taxon2:0.24)Clade2:0.42,taxon3:0.7)Clade1:0.36,((taxon4:0.35,taxon5:0.55)Clades4:0.250,taxon6:0.31)Clade3:0.8);
[200]"no_motif" #b1.5# {2,0.03}(((taxon1:0.32,taxon2:0.24)Clade2:0.42,taxon3:0.7)Clade1:0.36,((taxon4:0.35,taxon5:0.55)Clades4:0.250,taxon6:0.31)Clade3:0.8);
```
IsgGui: A New tool for Sequence simulation presented by Catherine Anderson

IsgGui: a graphical user interface for iSGv2

- Developed by Cate Anderson
- Programmed in java SE 6
- Works on top of installed indel-seq-gen-2.1
- Packaged in an executable jar file with needed libraries
- Set up entails the selection of various directories
Advantages

- Allows for quick addition of both global and guide tree parameters.
- Provides parameter format error checking.
- Provides parameter compatibility checking.
- Displays command-line text from GUI.
- Generate graphical representation of results.
Main Window
Go to live demo
Lineage File

```plaintext
MOTIFS =
{
    root:
        MARKER=b;
        NAME=PS00213: Lipocalin signature;
        PATTERN=[DEN]-{A}->{DENQGSTARK}x(0,2)=[DENQARK]-[LIVFY]
                   -{CP}G-{C}W-[FYWLRH]x{A}->[LIVMTA];

    root:
        MARKER=7;
        NAME=lipo_ma: Lipocalin partial template;
        PATTERN=x(5,20)-x(10,30);
}
```
iSGv2: indel-seq-gen (5)

Root sequence multiple alignment file

```
Taxon1
EGP-LDT-----AVEEE--QISGNWLGMRSTYHV--VS--L-FNQGILQ-----EV
Taxon2
ERS-LEMT---VVGDSGDAIDGVWLFQYINEVGLRQRLKF-----AALGTIKL
Taxon3
DGP-IDTIQTNVLDSS--DIAGRWYMILGDAMFRRTMKGL-----KNVLSGPL
Taxon4
PFPNLATKLIGQPDQ--EIIQGYELSHHSKSAIFGDTSM-----KLV-TK--
PS00213
***************bbbfbbbbfbbbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbbbfbb
4 components of the Main Window

- Basic parameters - entering basic global parameter.
- Advanced parameters - entering more specific controls
- Edit guide tree - allows additional partitions to be added or deleted from guide tree file
- Edit lineage file - allows changes to subtree (clade) parameters and motifs
Implementation of IsgGui (5)

3 display options

- Alignment display -
  - Can display any MSA in fasta format
  - Can display events within in alignment
  - Provides facility to compare alternate alignments

- Phylogeny display
  - Allows for the display of any tree in Newick format
  - Allows for the display of indel events on phylogeny
  - Allows for the display and editing of guide trees for iSGv2.

- Pixel-plot display
  - Allows for a larger area view of alignment
  - Allows for the comparison of up to four alternate alignments
Uses for IsgGui

- For evolution simulation
  - Used to learn indel-seq-gen
  - Used to debug parameters
  - Used to debug support files

- For alignment comparisons
  - Compare two full alignments with highlighting to indicate conserved areas between alignments.
  - Compare up to 4 alignments in Pixel Plot format
Discussion

For alignment comparisons

Reference: aaa4_fasta.ma  Section columns: 521 – 535

Other Alignment: aaa4_fasta_mafft.ma  Range of corresponding symbols: 530 – 591
Future Work

Gui

- Multi-thread capability
- Faster image processing

Functionality

- Provide scoring of alternate MSAs based on “true” reference alignment
- Evaluate effect on Phylogeny for inconsistent alignment
- Identify the sequence of indel events that cause most difficulty, and adapt alignment method to “consider” these possibilities
References - Alignment programs and curated alignments

- “CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice”

- “MUSCLE: a multiple sequence alignment method with reduced time and space complexity.”

- “MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform”
  Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma and Takashi Miyata (2002)

- Manually curated MSA’s
  http://hem.fyristorg.com/acacia/alignments.htm
References - iSGv2 and IsgGui

- “indel-Seq-Gen: A New Protein Family Simulator Incorporating Domains, Motifs, and Indels”
  Cory L. Strope, Stephen D. Scott and Etsuko N. Moriyama (2009)

- “indel-Seq-Gen v2.0.5 Manual”
  Cory L. Strope, Kevin Abel, Stephen D. Scott and Etsuko N. Moriyama (2010)

- “IsgGui: A Graphical User interface to enhance the use of iSGv2”
  Catherine Anderson, Cory L. Strope, and Etsuko N. Moriyama (2010)

- “IsgGui 1.00 User’s manual”
Thank you for your attention!

Are there any questions?