Evolutionary Analysis of Presynaptic Genes

Dexter D. Hadley
dexter@mail.med.upenn.edu

Advisors
- Junhyong Kim, PhD
- Lyle Ungar, PhD
- Maja Bucan, PhD

Kim Lab Meeting
10/11/2005
Goals

- Investigate genes involved in neuronal communication
- Conduct comparative analysis to identify cis-regulatory elements
  - Identify evolutionary conserved elements
  - Elucidate biological properties of non-coding conserved elements
- ... via *in silico* approach
Synaptic Transmission

- Synapse
  - Pre-Synaptic Terminal
  - Synaptic Cleft
  - Post-Synaptic Terminal
- Ca\(^{2+}\)/ Na\(^{+}\) sensing
- Neurotransmitter release

Kandall et al 1991
Pre-Synaptic Proteins

- Coordinated Exocytosis
- Large number of genes (130)
- Several large gene families
  - Synaptotagmins
  - Syntaxins
  - VAMPs
  - RABs

Kandall et al 1991
Why Synaptic Transmission?

- Biological Perspective
  - Basic functional unit of neuronal communication
  - Best understood neurological processes
  - Many disease associations
    - Mental Retardation, Autism, Schizophrenia...

- Genomics Perspective
  - Manageable number of genes and gene interactions
  - Modular and hierarchical system
  - Closed dedicated system

- Computational Perspective
  - High quality sequence available
  - Expression patterns available
  - Well conserved across species
  - Many problems still remain
Overview

- Dataset Generation
  - SynapseDB
- Molecular Evolution
  - Protein Analysis
  - Comparative Analysis Pipeline
- RNA Analysis
  - Expression across mouse tissue
- Detailed Curation
  - Transcription Factor Binding Sites (TFBS)
  - Novel Genes and Transcripts
SynapseDB

Database and analysis of genes involved in synaptic transmission
There is a subset of brain-specific genes with low level of expression during embryogenesis and in other adult tissues.
Species Selection

  - Most finished sequence
  - Best annotation
  - Direct relevance to disease

- Rodent Orthologs
  - Mouse (NCBIM33, May 2004)
  - Rat (RGSC3, June 2003)
Genomic Region of Interest (gROI)

- Consider upstream and downstream to the next annotated \textit{cis-genes}
Comparative Analysis Pipeline

- SynapseDB
  Pre-Synaptic Genes

- Orthologue Selection

- gROI Selection

- Multiple Alignment

- 5% MCS Selection

- MCS Classification

- Annotated MCS
Comparative Analysis Pipeline

1. SynapseDB
   Pre-Synaptic Genes

2. Orthologue Selection

3. Protein Analysis

4. gROI Selection

5. Multiple Alignment

6. 5% MCS Selection

7. RNA Analysis

8. MCS Classification

9. Detailed Curation

10. Annotated MCS

11. Experimental Validation
Synonymous & Nonsynonymous Substitutions

- $d_S$ is the relative rate of synonymous mutations per synonymous site
- $d_N$ is the relative rate of non-synonymous mutations per non-synonymous site
- $\omega = \frac{d_N}{d_S}$
  - If $\omega = 1 \rightarrow$ neutral selection
  - If $\omega < 1 \rightarrow$ negative (purifying) selection
  - If $\omega > 1 \rightarrow$ positive (Darwinian) selection
Evolutionary Analysis of Proteins

Do presynaptic genes show unique patterns of evolution?

Methods

- Calculate $d_N/d_S$ ratios for presynaptic genes
- Calculate $d_N/d_S$ ratios for all genes in genome
- Compare distributions
Evolutionary Analysis of Proteins

Presynaptic genes are under greater purifying selection than random genes.

Genomic $d_N/d_S$
- $\mu = 0.41$
- $\sigma^2 = 0.25$

Synaptic $d_N/d_S$
- $\mu = 0.08$
- $\sigma^2 = 0.09$
Evolutionary Analysis of Proteins

- How do gene families involved in neurotransmission evolve?
- Do evolution rates differ among paralogous genes?
- Is there evidence for accelerated evolution in large gene families?

Methods
- Compare $d_N/d_S$ rates across large gene families
Evolutionary Analysis of Proteins

- Rabs show consistently stronger purifying selection
- ≥ 1 family member sows accelerated evolution
- Functional diversification
- Stx11 shows human-specific patterns
Protein Analysis Summary

- Presynaptic genes are under greater purifying selection than random genes
- Rabs show consistently stronger purifying selection than other large families
- At least one family member is undergoing accelerated evolution in large gene families
- Stx11 shows human-specific patterns of protein evolution
Comparative Analysis Pipeline

- **SynapseDB**
  - Pre-Synaptic Genes

- **Orthologue Selection**

- **gROI Selection**

- **Multiple Alignment**

- **5% MCS Selection**

- **RNA Analysis**

- **MCS Classification**

- **Detailed Curation**

- **Annotated MCS**

- **Protein Analysis**

- **Experimental Validation**
Multi-species Conservation Scores

- Select most evolutionary conserved regions
  - From multiple alignment of sequences
- Defined as contiguous segment of bases exceeding a conservation threshold
  - Threshold derived from large genomic alignments
  - $\approx 5\%$ reference sequence selected

Thomas et al 2003
Comparative Sequence Analysis

Upstream
\[ \Sigma = 19\text{Mb} \]
\[ M = 64\text{Kb} \]
\[ \sigma^2 = 182\text{Kb} \]

Syn Gene
\[ \Sigma = 19\text{Mb} \]
\[ M = 49\text{Kb} \]
\[ \sigma^2 = 236\text{Kb} \]

Downstream
\[ \Sigma = 24\text{Mb} \]
\[ M = 47\text{Kb} \]
\[ \sigma^2 = 372\text{Kb} \]

133 gROI

\[ \Sigma = 63\text{Mb} \]
\[ M = 234\text{Kb} \]
\[ \sigma^2 = 610\text{Kb} \]
MCS Classification

- **All MCS**
  - $61,000 +$
  - $\mu = 49$ bp
  - $\sigma^2 = 47$ bp
  - $\mu = 471$ MCS/gROI
  - $\sigma^2 = 676$ MCS/gROI

- **Large MCS (>300bp)**
  - $252$
  - $\mu = 2.4$ LMCS/gROI
  - $\sigma^2 = 4.7$ LMCS/gROI

---
MCS Similarity

- How much similarity do MCS share with each other?
- How much sharing is there between similarly typed MCS?
- How are shared MCS distributed in the genome?

Methods
- BLAST all MCS against themselves
MCS Similarity

Low level of sequence similarity among MCS

< 1% had a significant blast hit to another MCS
Most MCS that share similar sequence tend to be similarly typed.
Coding MCS Similarity

- Coding MCS show similarity among paralogs
Noncoding MCS tend to be duplicated in same gROI
Example Duplications

Cdh2 Intron

Cdh2 Upstream
MCS Conservation

- How conserved are the largest MCS in other species?
- Is there a difference between mammalian and non-mammalian conservation?

Methods
- BLAT all LMCS against additional genomes
MCS remarkably more conserved in mammals

UTR elements relatively highly conserved in mammals

UTR elements relatively less conserved in non-mammalian vertebrates

- Chicken is the exception where UTR conservation is most conserved in a non-mammalian vertebrate
Comparative Analysis Summary

- Low level of sequence similarity among MCS
  - MCS that share similar sequence tend to have similar locations relative to exons
  - Coding MCS show similarity among paralogs
  - Noncoding MCS tend to be duplicated locally
- MCS remarkably more conserved in mammals
- UTR elements relatively highly conserved in mammals
  - Chicken is the exception where UTR elements relatively more conserved in non-mammalian vertebrate
MCSs with significant secondary structure shows transcription in mouse tissues

- RT-PCR in showing transcription in 15 tissues
- Probes span three intronic (2b) MCSs and their neighboring exons
Comparative Analysis Pipeline

SynapseDB
Pre-Synaptic Genes

Orthologue Selection

Protein Analysis

gROI Selection

Multiple Alignment

5% MCS Selection

RNA Analysis

MCS Classification

Detailed Curation

Annotated MCS

Experimental Validation
Detailed Curation

- Do MCS represent known TFBS?
- Are different classes of MCS more likely to represent TFBS?

Method
- Calculate likelihood ratios that MCS span known TFBS
## MCS and TFBS

<table>
<thead>
<tr>
<th>Type</th>
<th>P(TFBS)</th>
<th>P(null)</th>
<th>Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>1b (partial exon)</td>
<td>0.34</td>
<td>0.030</td>
<td>11.24</td>
</tr>
<tr>
<td>1a (exonic)</td>
<td>0.29</td>
<td>0.029</td>
<td>9.96</td>
</tr>
<tr>
<td>2a (intergenic)</td>
<td>0.13</td>
<td>0.018</td>
<td>7.33</td>
</tr>
<tr>
<td>2b (intrinsic)</td>
<td>0.13</td>
<td>0.019</td>
<td>7.01</td>
</tr>
<tr>
<td>1c (UTRs)</td>
<td>0.19</td>
<td>0.031</td>
<td>6.08</td>
</tr>
<tr>
<td>All</td>
<td>0.14</td>
<td>0.019</td>
<td>7.27</td>
</tr>
</tbody>
</table>

- MCS likely to span TFBS
Detailed Curation

- Do some MCS hit novel genes or novel transcripts of known genes?

- Method
  - Use ESTs aligned to the genome and DoTS to select potentially misannotated noncoding MCS
Database of Transcribed Sequences (DoTS)

- ESTs and RNAs from GenBank
  - Sequence trimming
  - Repeat removal
  - Clustering
  - Assembling
  - Annotation
  - Genomic alignment
- Output DoTS Transcripts (DTs) and Genes (DGs).
Detailed Curation Results

- Novel Genes: 56
  - Nested Anti-Sense: 29
  - Intergenic: 27
- Novel Transcripts: 112
Detailed Curation Examples

- Novel nested antisense gene in NLGN1 intron

- Novel gene downstream PCLO
Detailed Curation Summary

- Coding potential useful to identify novel coding ORFS
- Aligned ESTs together with DoTS can identify putative novel genes and transcripts
Acknowledgements

- Advisors:
  - Junhyong Kim, PhD
  - Lyle Ungar, PhD
  - Maja Bucan, PhD
- Tara Murphy
- Otto Valladares
- Bucan and Kim Labs