Software Prediction of the Effects of Single Nucleotide Polymorphisms

Angela Collie
CSE 527
December 13, 2004
Objective

Examine the ability of two web-based programs to predict the effect of a single nucleotide polymorphism on a protein.
Single Nucleotide Polymorphisms (SNPs)

- 99.9% of the 3.2 billion base pairs in the human genome are the same.
- SNPs are single base pair changes and account for much of the variation.
  - Minor allele is defined as present in >1% of the population.
  - “Common” alleles are present in >10% of the population.
  - There are approximately 11 million SNPs in the genome, corresponding to 1 base pair change every 300 bases.

- Haplotype

Nonsynonymous SNPs

- nsSNPs are SNPs that are present in the coding region of a gene and result in an amino acid change in the resulting protein.
  - This can affect the 3D structure or interactions with other proteins.
- SNPs in the promoter and exons of a gene are thought to be the most harmful to a protein.

dbSNP

- Public database maintained by the NCBI.
- Must recent build had over 10 million SNPs. 5 million have been validated.
- Data is linked to gene and other NCBI databases, including 3D structure representations for some SNPs.

http://www.genome.utah.edu/genesnps/cgi-bin/frame.cgi?gene_id=440
Polymorphism Phenotyping (PolyPhen)

- Web-based tool for predicting the effect of a nsSNP on a protein.
- Utilizes a combination of 3D structural parameters and sequence homology to make prediction based on rules.
- Input is protein sequence (or ID #) and position of amino acid substitution and amino acid variants.
- Returns predictions of “probably damaging,” possibly damaging,” “benign,” and “unknown.”

http://www.bork.embl-heidelberg.de/PolyPhen/

PolyPhen Algorithm: Step 1

- Characterization of substitution site.
- Checks protein database for protein features.
- Uses several program add-ins to identify transmembrane, coil and signal peptide regions.
- If substitution is in a transmembrane region, a score is calculated to determine effect.

http://www.bork.embl-heidelberg.de/PolyPhen/

PolyPhen Algorithm: Step 2

- PolyPhen uses BLAST against a protein database to identify sequences with 30-94% homology to input sequence.
- Position-Specific Independent Counts (PSIC) is run. This returns a score that is based on the log likelihood ratio of the amino acid \(a\) occurring at position \(i\) compared to the background frequency of amino acid \(a\).
- Ratio is corrected to account for the limited number of sequences available and the interdependence of sequences.

\[
W(\alpha,i) = \ln \left( \frac{p(\alpha,i)}{q_\alpha} \right)
\]

Sequence homology scores are based on this log likelihood ratio.

http://www.bork.embl-heidelberg.de/PolyPhen/

PolyPhen Algorithm: Step 3

- PolyPhen BLASTs the sequence against the user-chosen PDB or PQS databases to find proteins of sequence identity of at least 50%.
- Several structural parameters are then calculated using a protein database and another add-in.
- Polyphen then checks contacts of the variant amino acid with ligands, interactions between parts of the protein, and critical residues.

http://www.bork.embl-heidelberg.de/PolyPhen/

## PolyPhen Algorithm: Rules

<table>
<thead>
<tr>
<th>PSIC score difference</th>
<th>Substitution site properties</th>
<th>Substitution type properties</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arbitrary</td>
<td>functional or bond formation site</td>
<td>Arbitrary</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>Not considered</td>
<td>In a region annotated or predicted as transmembrane</td>
<td>PHAT matrix difference resulting from substitution is negative</td>
<td>Possibly damaging</td>
</tr>
<tr>
<td>Less than 0.5</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td>Benign</td>
</tr>
<tr>
<td>Greater than 1.0</td>
<td>Atoms are closer than 3.0 Å to atoms of a ligand or residue annotated as BINDING, ACT_SITE, LIPID, METAL</td>
<td>Arbitrary Absolute change of accessible surface propensity is 0.75 or absolute change of side chain volume is 60 Absolute change of Normed accessibility ACC 5%</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>Between 0.5 and 1.5</td>
<td>Nominal accessibility ACC 15%</td>
<td>accessible surface propensity is 1.0 or absolute change of side chain volume is 80</td>
<td>Possibly damaging</td>
</tr>
<tr>
<td>Between 0.5 and 1.5</td>
<td>Nominal accessibility ACC 5%</td>
<td>accessible surface propensity is 1.0 or absolute change of side chain volume is 80</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>2.0</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td>Possibly damaging</td>
</tr>
<tr>
<td>Greater than 2.0</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td>Probably damaging</td>
</tr>
</tbody>
</table>

Table legend: One row corresponds to one rule, which may consist of several parts connected by logical AND. If no evidence for a damaging effect is seen, substitution is considered benign.

http://www.bork.embl-heidelberg.de/PolyPhen/
This variant is predicted to be possibly damaging.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Available data</th>
<th>Prediction basis</th>
<th>Substitution effect</th>
<th>Prediction data</th>
</tr>
</thead>
<tbody>
<tr>
<td>possibly damaging</td>
<td>FT alignment structure</td>
<td>Structure</td>
<td>1.1.1: structural effect, buried site, hydrophobicity</td>
<td>PSIC score difference: 0.839 normed accessibility: 0.14 hydrophobicity change: 1.07</td>
</tr>
</tbody>
</table>

Remarks
Closest contact with other chains: GLN 125C, distance 2.800 Å

Details
PSIC PROFILE SCORES FOR TWO AMINO ACID VARIANTS

<table>
<thead>
<tr>
<th>Score 1</th>
<th>Score 2</th>
<th></th>
<th>Score1-Score2</th>
<th>Observations</th>
<th>Diagnostics</th>
<th>Multiple alignment around substitution position</th>
</tr>
</thead>
<tbody>
<tr>
<td>+0.945</td>
<td>+0.106</td>
<td>0.839</td>
<td></td>
<td>90</td>
<td>cached</td>
<td>Sequences: Flanks:</td>
</tr>
</tbody>
</table>

http://www.bork.embl-heidelberg.de/PolyPhen/
Sorting Intolerant From Tolerant (SIFT)

• Web-based tool for predicting the effect of a nsSNP on a protein.
• Utilizes sequence homology to predict effect.
• Aligned protein sequences are from BLink.
• Input is GI# (unique for protein) or protein sequence. SNP amino acid substitutions and position can also be submitted.
• Returns predictions of “affect protein function” and “tolerated” for each SNP. Also returns normalized score and median sequence information.

SIFT Algorithm

\[ p_{ca} = \frac{N_c}{(N_c + B_c)} \cdot g_{ca} + \frac{B_c}{(N_c + B_c)} \cdot f_{ca} \]

• “\( p_{ca} \), the probability of amino acid \( a \) at position \( c \), is a weighted average of the observed amino acid frequencies in the alignment and the estimated unobserved frequencies.”
  – \( N_c \) is the number of sequences at position \( c \).
  – \( B_c \) is an exponential function that returns the number of pseudocounts based on amino acid frequencies in a predetermined matrix.
  – \( g_{ca} \) is a sequence weighted frequency that \( a \) appears at \( c \) in the alignment.
  – \( f_{ca} \) is a frequency of pseudocounts.
• Normalization.

Ng PC, Henikoff S. Genome Res. 2001 May;11(5):863-74.
http://blocks.fhcrc.org/sift/SIFT.html
Sample SIFT Output

• Threshold for tolerance is 0.5.
• For position 1M
  – Predict not tolerated: ywvtspqnkhihgfedca
  – Predict tolerated: M
  – Normalized probabilities for each amino acid can also be obtained
• For substitution: A94T
  – Substitution at pos 94 from A to T is predicted to AFFECT PROTEIN FUNCTION with a score of 0.02.
  – Median sequence conservation: 2.64
  – Sequences represented at this position: 48

Ng PC, Henikoff S. Genome Res. 2001 May;11(5):863-74. 
http://blocks.fhcrc.org/sift/SIFT.html
Genes of Interest

IL-1R1
IRAK1
IRAK4

TNF
IL-6
### PolyPhen AND SIFT Results

<table>
<thead>
<tr>
<th>Gene</th>
<th>Reference aa residue / SNP aa residue</th>
<th>PolyPhen prediction</th>
<th>SIFT prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL1R1</td>
<td>Ala [A]/Gly [G]</td>
<td>Benign</td>
<td>Tolerated</td>
</tr>
<tr>
<td>IRAK1</td>
<td>Ser [S]/Leu [L]</td>
<td>Possibly damaging</td>
<td>Tolerated</td>
</tr>
<tr>
<td></td>
<td>Arg [R]/Gly [G]</td>
<td>Possibly damaging</td>
<td>Affect Protein Function</td>
</tr>
<tr>
<td></td>
<td>Cys [C]/Ser [S]</td>
<td>Probably damaging</td>
<td>Tolerated</td>
</tr>
<tr>
<td></td>
<td>Phe [F]/Ser [S]</td>
<td>Benign</td>
<td>Affect Protein Function</td>
</tr>
<tr>
<td></td>
<td>Arg [R]/His [H]</td>
<td>Benign</td>
<td>Tolerated</td>
</tr>
<tr>
<td></td>
<td>Thr[T]/Ile [I]</td>
<td>Benign</td>
<td>Tolerated</td>
</tr>
<tr>
<td>IRAK4</td>
<td>Ser [S]/Arg [R]</td>
<td>Benign</td>
<td>Affect Protein Function</td>
</tr>
<tr>
<td></td>
<td>His [H]/Arg [R]</td>
<td>Benign</td>
<td>Tolerated</td>
</tr>
<tr>
<td></td>
<td>Ala [A]/Thr [T]</td>
<td>Benign</td>
<td>Tolerated</td>
</tr>
<tr>
<td>TNF</td>
<td>His[H]/Asn [N]</td>
<td>Benign</td>
<td>Tolerated</td>
</tr>
<tr>
<td></td>
<td>Pro [P]/Leu [L]</td>
<td>Possibly damaging</td>
<td>Tolerated</td>
</tr>
<tr>
<td></td>
<td>Ala [A]/Thr [T]</td>
<td>Benign</td>
<td>Affect Protein Function</td>
</tr>
<tr>
<td></td>
<td>Ile [I]/Asn [N]</td>
<td>Probably damaging</td>
<td>Affect Protein Function</td>
</tr>
<tr>
<td>IL6</td>
<td>Pro [P]/Ser [S]</td>
<td>Benign</td>
<td>Tolerated</td>
</tr>
<tr>
<td></td>
<td>Leu [L]/Pro [P]</td>
<td>Probably damaging</td>
<td>Affect Protein Function</td>
</tr>
<tr>
<td></td>
<td>Asp [D]/Val [V]</td>
<td>Benign</td>
<td>Tolerated</td>
</tr>
<tr>
<td></td>
<td>Asp [D]/Glu [E]</td>
<td>Benign</td>
<td>Tolerated</td>
</tr>
</tbody>
</table>

IRAK4 is recently described so there may not have been enough sequences for prediction.
Software Comparison

- 18 SNPs were examined in 5 genes.
- PolyPhen and SIFT have different predictions for several SNPs.
- None of these SNPs have been described in OMIM or in the literature.
Conclusion

- Two web-based software programs were used to predict the effect of 18 SNPs on 5 genes in the IL-1B signaling pathway.
- Software predictions must be verified with experimental data.
- Predictions will improve with additional homologous sequences and three-dimensional structure.