V. Predict effect of sequence variation

short title: pp2_sav2

lecture: Protein Prediction 2 (for Computational Biology) - Protein function
TUM winter semester
Videos: YouTube / www.rostlab.org

THANKS:

Dmitrij Nechaev

Special lectures:
• (TBC)

No lecture:
• 10/30 no lecture
• 11/01 All Saints
• 11/13 SVV (student rep)
• 11/22 Thanksgiving
• 12/06 Dies Academicus (TUM)
• 12/20-01/06 - no lecture Xmas+
• 01/08 no lecture?

LAST lecture: Jan 24 (followed by 2 wrap-up sessions)

Examen: Feb 07 10:00-13:00, LMU physics

Makeup: TBC: Apr 23 & Apr 25, 2019 - lecture time
V. Predict effect of mutations
V.1 SNP|SNV/SBV effect: Meaning 2
Similar sequence -> similar structure/function!

C Sander & R Schneider 1991 Proteins 9:56-68
B Rost 1999 Prot Engin 12:85-94
Types of sequence variation - DNA/RNA level

- **Substitutions**
  - A single nucleotide change, e.g. A -> C

- **Insertions and Deletions**
  - Addition or removal of one or more nucleotides
  - “Indels” may be just that but can also denote a special case of both an insertion and a deletion simultaneously

- **Duplications**
  - Repeat of one or more nucleotides

- **Inversion**
  - More than one nucleotide replaces its reverse complement
  - For example: ...AGGCTGATT... to ...AGGT*CAGTT...
Types of sequence variation - Protein level

- Looking at variation only in the coding region:
  - Substitutions
    - Missense, non-synonymous
      - Variant leads to a change in the resulting amino acid
    - Silent, synonymous
      - Variant does not change the resulting amino acid
    - Nonsense
      - Variant leads to a premature stop codon
  - Insertion, Deletion, Duplication
    - Same idea as on the DNA level
  - Inversions?
    - AAUAGA (Asn Arg) to UCUAUU (Ser Ile)
Standardization efforts for example as the Sequence Variant Nomenclature

Single Nucleotide Variant vs. Single Nucleotide Polymorphism
- Often used interchangeably
- Traditionally: SNP is a position where each variant occurs in > 1% of the population

“Mutation”
- A change, possibly disease-causing, not well defined

Single amino acid variant (SAV)
- 19 non-native
  - Substituting the native amino acid by all others
  - In contrast to SNV-possible
### Sequence variation - Some more nomenclature

**SAV-possible**

- Substitute native amino acid by all reachable by SNV
- e.g. UGG(W)→UUU | UUC(F) not SAV-possible (2 substitutions)

**UGG (Trp/W) Tryptophan**

<table>
<thead>
<tr>
<th>1st base</th>
<th>2nd base</th>
<th>3rd base</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>UUU</td>
<td>U</td>
</tr>
<tr>
<td></td>
<td>(Phe/F) Phenylalanine</td>
<td>(Ser/S) Serine</td>
</tr>
<tr>
<td>U</td>
<td>UUC</td>
<td>C</td>
</tr>
<tr>
<td>U</td>
<td>UUA</td>
<td>U</td>
</tr>
<tr>
<td>U</td>
<td>UUG</td>
<td>U</td>
</tr>
<tr>
<td>C</td>
<td>CUU</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>(Leu/L) Leucine</td>
<td>(Pro/P) Proline</td>
</tr>
<tr>
<td>C</td>
<td>CUC</td>
<td>C</td>
</tr>
<tr>
<td>C</td>
<td>CUA</td>
<td>C</td>
</tr>
<tr>
<td>C</td>
<td>CUG</td>
<td>C</td>
</tr>
<tr>
<td>A</td>
<td>AUU</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>(Ile/I) Isoleucine</td>
<td>(Thr/T) Threonine</td>
</tr>
<tr>
<td>A</td>
<td>AUC</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>AUA</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>AUG[A]</td>
<td>A</td>
</tr>
<tr>
<td>G</td>
<td>GUU</td>
<td>G</td>
</tr>
<tr>
<td></td>
<td>(Val/V) Valine</td>
<td>(Ala/A) Alanine</td>
</tr>
<tr>
<td>G</td>
<td>GUC</td>
<td>G</td>
</tr>
<tr>
<td>G</td>
<td>GUA</td>
<td>G</td>
</tr>
<tr>
<td>G</td>
<td>GUG</td>
<td>G</td>
</tr>
</tbody>
</table>
Example of the effects of sequence variation

- nsSNP: Sickle-cell anaemia
- Hemoglobin subunit beta (HBB_HUMAN)
  - GAG to GTG, p.Glu7Val
- Hemoglobin A aggregates upon deoxygenation
  - Red blood cells become sickle shaped
    - Cells are less stable, and get stuck in capillaries
- Heterozygotes
  - Mostly no adverse effects (unless at severe conditions such as high altitude)
  - Decreases risk of Malaria infection!
How can synonymous SNVs have an effect?
Example of the effects of sequence variation

- **Synonymous SNV: Crouzon Syndrome**
  - Variants in Fibroblast growth factor receptor 2 (**FGFR2_HUMAN**)
  - One synonymous variant leads to an upstream splice site usage and exon truncation

General mechanisms:
- RNA regulation and post-transcriptional processing
- Translation speed (**cf.** co-translational folding)
- mRNA structure and stability
  - May degrade faster or it may be harder to initiate translation

[Diagram showing RNA regulation and mRNA structure with examples of normal and Crouzon splice sites and nucleotide changes](image-url)
Example of the effects of sequence variation

Deletion: Cystic Fibrosis

- Cystic fibrosis transmembrane conductance regulator (CFTR HUMAN)
  - p.Phe508del
    - CFTR misfolds and is not transported out of the ER
    - Lack of ion transport across cell membrane
Types of sequence variation - Larger changes

- **Structural variants**
  - Change on the level of 1kb to 1mb (10^3 to 10^6 base pairs)
  - Insertions, Deletions, Inversions
  - Copy Number Variation
    - Sections of the genome are repeated and the number of repeats differs between individuals
    - Caused by duplications and deletions
    - “Repeats” can be anything from di-nucleotides to whole genes

- **Chromosomal rearrangements**
  - Even larger changes. Often not clearly distinguished from structural variants

1. Deletion
2. Duplication
3. Inversion

Insertion

**Additional images:***
- [Two Chromosome Mutations](https://en.wikipedia.org/wiki/File:Two_Chromosome_Mutations.png)
Copy number variation: Huntington’s disease

- **Huntingtin** (HD_HUMAN)
- Poly-Gln repeats are highly polymorphic in population (~10-35 in healthy individuals)
- Short repeats lend themselves to errors in DNA replication
- 36 and more repeats lead to apoptosis of neurons due to aggregation
- Interestingly number of repeats correlates with age of disease onset
Example of the effects of sequence variation

Chromosome duplication / translocation: Trisomy 21

- 3 copies of chr21 through errors during meiosis
- Also 2 copies and some more due to a translocation in a parent
V.2 SAV effect: 1st generation methods
Similar sequence -> similar structure/function!

C Sander & R Schneider 1991 Proteins 9:56-68
B Rost 1999 Prot Engin 12:85-94
SIFT: Sorting Intolerant From Tolerant
Steven Henikoff

Shapers and Shakers

Fred Hutchinson Cancer Center, Seattle
HHMI (Howard Hughes Medical Institute)

papers:
- >300 papers (Nov 2011)
- 3 >1,000 citations (end 2011)
- 72 over 100
- H-index 83 (ISI Nov 2011)

Paradigm changes
- gene in gene - in intron (1986)
- histones NOT only in octamers (2004)
- DNA-methylation in histones: H2.AZ in histone spool promotes gene expression (2008): NOT DNA-methylation shuts off genes (important for cancer drug development)
Pauline Crystal Ng

- Genome Institute of Singapore/ASTAR (Senior Investigator)

CV:
- BS: Caltech
- PhD (Bioengineering): Univ of Washington, Fred Hutchinson, Seattle (with Steven Henikoff)
- Postdoc: Illumina
- Assistant Professor: J Craig Venter Institute, Seattle
- Sr Investigator POLARIS / ASTAR / Gen Inst Singapore

papers:
- >20 papers (Nov 2011)
- 3 >300 citations (end 2011)
- common name: >300 papers, most of the top cited ones from HER!
Test Set: 3 proteins, ~6500 mutants

PolyPhen:
Polymorphism
Phenotyping
Shamil R Sunyaev

- Harvard Medical School  
  (Assistant Professor)
- CV:
  - PhD: Moscow Inst. of Physics & Technology EMBL
  - Postdoc: EMBL
- papers:
  - >50 papers (Nov 2011)
  - 3 >100 citations (end 2011)
PolyPhen

Test Set: 1551 SWISS-PROT & 440 cross-species variants

Train/Test Set: 3768 HGMD variants and 2309 cross-species variants

V.3 SAV effect: Data
Misfunction/neutral
SNAP data set: PMD

Y Bromberg & B Rost 2007 NAR 35:3823-35

ENTRY A000006 - Variant 2616650
AUTHORS Farooqi I.S., Yeo G.S.H., Keogh J.M., Aminian S., Jebb S.A., Butler G., Cheetham T. & O'Rahilly S.
CROSS-REFERENCE MC4R_HUMAN
PROTEIN Melanocortin 4 receptor (MC4R);
CHANGE-POINT Asn 62 Ser (homozygous)
DISEASE In obesity
FUNCTION Responsiveness to alpha-MSH [-]

T Kawabata, M Ota & K Nishikawa (1999)
The protein mutant database. NAR 27: 355-7
SNAP data set: PMD

ENTRY A000006 - Variant 2616650
AUTHORS Farooqi I.S., Yeo G.S.H., Keogh J.M., Aminian S., Jebb S.A., Butler G., Cheetham T. & O'Rahilly S.
CROSS-REFERENCE MC4R_HUMAN
PROTEIN Melanocortin 4 receptor (MC4R);
CHANGE-POINT Asn 62 Ser (homozygous)
DISEASE In obesity
FUNCTION Responsiveness to alpha-MSH [-]

Effect: 40,641
Neutral: 14,334


Machine-Learning handles imbalance?

- 74% effect
- 26% neutral
SNAP data: neutral

EC# = general_class . acts_on_class . further_class_spec . spec_by_substrate_class

3.1.3.48 → hydrolase . on ester bonds . phosphoric monoester cmpnds . PTP-phosphotase

Same EC# = Same Function

Query: 61 YYQLFELMNKVGAFSHLRLEHTHTFVNKGRTGALDFRFFTGAPFNGLKAFFTTSQSL 120
YY LF LM KVGA +LRLKEHTHTFVN+GGR G LDFRF TGAPFNGLKAFFTTSQQL

Sbjct: 61 YYNLFNLMKVGAKQNLRLKEHTHTFVNQGGRIGELDFRFLTGAPFNGLKAFFTTSQLDT 120

Neutral: 26,840


### Availability of data for nsSNP limiting

<table>
<thead>
<tr>
<th></th>
<th>Effect</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMD</td>
<td>40,641</td>
<td>14,334</td>
</tr>
<tr>
<td>EC</td>
<td></td>
<td>26,840</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>Effect:</strong></td>
<td><strong>Neutral:</strong></td>
</tr>
<tr>
<td></td>
<td>40,641</td>
<td>41,174</td>
</tr>
</tbody>
</table>

- **81,815 variants**
- **6,821 proteins**

# SNAP data: by accessibility

<table>
<thead>
<tr>
<th></th>
<th>Proteins</th>
<th>Effect</th>
<th>Neutral</th>
<th>Total</th>
<th>Ratio neutral to effect</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>6,413</td>
<td>39,987</td>
<td>40,830</td>
<td>80,817</td>
<td>1.02</td>
</tr>
<tr>
<td><strong>Buried</strong></td>
<td>5,144</td>
<td>19,741</td>
<td>14,800</td>
<td>34,541</td>
<td>0.75</td>
</tr>
<tr>
<td><strong>Intermediate</strong></td>
<td>4,841</td>
<td>12,285</td>
<td>13,073</td>
<td>25,358</td>
<td>1.06</td>
</tr>
<tr>
<td><strong>Exposed</strong></td>
<td>4,150</td>
<td>7,961</td>
<td>12,957</td>
<td>20,918</td>
<td>1.62</td>
</tr>
</tbody>
</table>

**Protein exposure categories:**
- **Exposed:** 49% buried, 36% exposed
- **Intermediate:** 49% buried, 32% exposed
- **Buried:** 43% neutral, 57% effect

**Effect distribution:**
- **Neutral:** 32%
- **Effect:** 68%

**Neutral to total ratio:**
- Total: 1.02
- Buried: 0.75
- Intermediate: 1.06
- Exposed: 1.62

---

V.4 SAV effect: 2\textsuperscript{nd} generation methods
Classification of SAV-effect prediction methods

1st generation:
use simple small data set of variants only: SIFT, PolyPhen, SNPs3D

2nd generation:
accumulated larger data sets, advent of machine learning
e.g. SNAP, SIFT2, PolyPhen-2

3rd generation:
large data sets, including diversity of information, essentially machine learning, only
SNAP - Yana Bromberg (now Rutgers University*) AND TUM-IAS fellow

...SNAP...
Some other methods in a jiffy
SDM (Site Directed Mutator)

Thermodynamic cycle

\[
\begin{align*}
\text{WT}^U_j & \xrightarrow{\Delta G_{\text{wt-mut}}^U} \text{Mut}^U_k \\
\Delta G_{\text{wt}}^{U-F} & \xrightarrow{} \Delta G_{\text{mut}}^{U-F} \\
\text{WT}^F_i & \xrightarrow{\Delta G_{\text{wt-mut}}^F} \text{Mut}^F_k
\end{align*}
\]

CM Topham, N Srinivasan & TL Blundell (1997) Protein Eng. 10, 7-21

© http://mordred.bioc.cam.ac.uk/~sdm/sdm_theory.php

© Burkhard Rost

© BBC

Professor Sir Tom L Blundell

© Enrico Coen
SDM (Site Directed Mutator)

Thermodynamic cycle

CM Topham, N Srinivasan & TL Blundell (1997) Protein Eng. 10, 7-21

© http://mordred.bioc.cam.ac.uk/~sdm/sdm_theory.php
Changes in binding energy learned by neural networks from statistical potentials

Y Dehouck et al & M Rooman (2009) Bioinformatics 25, 2537-43

© http://babylone.ulb.ac.be/poppmusic
I-Mutant-2 - Prediction of Protein Mutant Stability Changes

Changes in binding energy learned by SVM from ProTherm

Emidio Capriotti, Piero Fariselli & Rita Casadio (2005) NAR 33, W306-10

© [link](http://folding.biofold.org/i-mutant/i-mutant2.0.html)
AutoMute - AUTOmated server for predicting ... functional consequences of amino acid MUTations in protEins

Changes in binding energy learned by Machine Learning considering 3D environment


©http://proteins.gmu.edu/automute/
PROVEAN - protein variation effect analyzer

- **Extension of the SIFT idea**
  - Conservation is scored “locally” for all clusters of homologs
  - Cluster are then weighted equally for the final score

- **Scores coding SNVs, insertions, deletions**

\[
\Delta SW(A, \text{var}, B) = SW(A_{\text{var}}, B) - SW(A, B)
\]

\[
\text{PROVEAN score} = \frac{1}{N} \sum_{c=1}^{N} \left( \frac{1}{N_c} \sum_{i=1}^{N_c} \Delta_{c,i} \right),
\]

Yongwok Choi, GE Sims, S Murphy, JR Miller, Agnes P Chan
Mutation Taster

- Bayesian network
- OMIM disease vs. 1KG

ACTION
“Evolutionary Action”

63 annotations from genomic databases combined in an SVM

- Conservation
- Regulatory information (e.g. transcription factor binding)
- Transcript information (e.g. exon/intron boundaries)
- “Protein-level” scores (e.g. SIFT, PolyPhen predictions)

Scores SNVs and insertions and deletions in coding and non-coding regions

For Training

- Variants without effect are all that differ between human and a human-chimp common ancestor and appear in large numbers in the human population
- All other in silico simulated variants are deleterious


Slide adapted from Michael Bernhofer & Jonas Reeb

© Burkhard Rost ROSTLAB
MutationAssessor

**Question**
- Which mutations are functional?
- Which mutations have implications for cancer progression?

**Hypothesis**
- Mutations in evolutionarily conserved residues are likely functional
- Mutations in non-conserved residues are likely neutral
- Analysis of evolutionary conservation patterns can discriminate between functional and non-functional mutations

**Method**
- Conserved residues: conserved across entire family
- Specificity residues: conserved within subfamily, vary between subfamilies

<table>
<thead>
<tr>
<th>Protein</th>
<th>Conservation Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGFR_HUMAN</td>
<td>GLKELPMRNLOEILHGAVRFSN</td>
</tr>
<tr>
<td>Q8M18_PIG</td>
<td>GLRELPNRLSEILNGVQSNN</td>
</tr>
<tr>
<td>EGFR_CHICK</td>
<td>GLRELPKRLSEILNGVQSNN</td>
</tr>
<tr>
<td>ERBB4_HUMAN</td>
<td>GLQELGKLNLEILNGGYYVDQN</td>
</tr>
<tr>
<td>INSR_MOUSE</td>
<td>HLKELGLNLMNITGRSVRIEKN</td>
</tr>
<tr>
<td>ILPR_BRALA</td>
<td>DMQKIGLSTQLNITGRSVRIEKN</td>
</tr>
<tr>
<td>IGFI1_XENLA</td>
<td>DLKEIGLNLNITGRSVRIEKN</td>
</tr>
</tbody>
</table>

**Functional Impact Score** = conservation score + specificity score

**Validation**
- Disease-associated polymorphisms

**Functional impact: disease or neutral?**
- 80% classification accuracy in separation of 36K common polymorphisms (assumed neutral) from 19K disease-associated variants (assumed functional)
- AUC = 0.86

**Functional impact in cancer: stronger or weaker?**
- 10K point non-synonymous mutations in COSMIC.v49:
  - non-recurrent (observed in only one sample) vs recurrent (observed in 2 or more samples): 69% classification accuracy, AUC=0.75
  - non-recurrent vs highly recurrent (observed in 5 or more samples): 78% classification accuracy, AUC=0.84

**Public server**
- One-stop shop for protein mutation analysis
- Rich annotations, pathways, 3D structure, binding sites, etc.
- WEBAPI allows batch submission, querying mutation functional impact score and all annotations, linking to mutation views in MSA / 3D

B Reva, Y Antipin & C Sander (2011) NAR 39:e118
Evolutionary Couplings

easy

inverse problem

© Burkhard Rost


T Hopf, C Sander, D Marks (2016) submitted
Evolutionary Couplings

To what extent do we see a pair of amino acids more/less often than expected by chance?

\[ f_{ij}(A_i, A_j) - f_i(A_i)f_j(A_j) \]

single column frequencies: \( f_i(A_i) \)

column pair frequencies: \( f_{ij}(A_i, A_j) \)
Evolutionary Couplings predict effect of variants
How to compare methods?
Prediction Performance Comparison

*SDM, Crescendo, (PICCOLO, CREDO and BIPA), SIFT, MUpro, MAPP and I-Mutant2.0*

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
<th>Sum</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDM</td>
<td>535</td>
<td>274</td>
<td>4594</td>
<td>3025</td>
<td>8428</td>
<td>15.03</td>
<td>94.37</td>
<td>60.86</td>
</tr>
<tr>
<td>Crescendo</td>
<td>186</td>
<td>126</td>
<td>4482</td>
<td>3695</td>
<td>8489</td>
<td>4.79</td>
<td>97.27</td>
<td>54.99</td>
</tr>
<tr>
<td>PICCOLO</td>
<td>220</td>
<td>257</td>
<td>4920</td>
<td>3746</td>
<td>9143</td>
<td>5.55</td>
<td>95.04</td>
<td>56.22</td>
</tr>
<tr>
<td>CREDO</td>
<td>380</td>
<td>413</td>
<td>4764</td>
<td>3586</td>
<td>9143</td>
<td>9.58</td>
<td>92.02</td>
<td>56.26</td>
</tr>
<tr>
<td>BIPA</td>
<td>80</td>
<td>20</td>
<td>5157</td>
<td>3886</td>
<td>9143</td>
<td>2.02</td>
<td>99.61</td>
<td>57.28</td>
</tr>
<tr>
<td>COMBINED</td>
<td>1252</td>
<td>984</td>
<td>4193</td>
<td>2714</td>
<td>9143</td>
<td>31.57</td>
<td>80.99</td>
<td>59.55</td>
</tr>
<tr>
<td>SIFT</td>
<td>2709</td>
<td>2071</td>
<td>3011</td>
<td>1092</td>
<td>8883</td>
<td>71.27</td>
<td>59.25</td>
<td>64.39</td>
</tr>
<tr>
<td>MAPP</td>
<td>2659</td>
<td>1642</td>
<td>2395</td>
<td>1065</td>
<td>7761</td>
<td>71.40</td>
<td>59.33</td>
<td>65.12</td>
</tr>
<tr>
<td>I-Mutant2.0</td>
<td>1485</td>
<td>1677</td>
<td>2189</td>
<td>1061</td>
<td>6412</td>
<td>58.33</td>
<td>56.62</td>
<td>57.30</td>
</tr>
<tr>
<td>MUpro</td>
<td>175</td>
<td>146</td>
<td>5031</td>
<td>3791</td>
<td>9143</td>
<td>4.41</td>
<td>97.18</td>
<td>56.94</td>
</tr>
</tbody>
</table>

Table 1. TP= True Positives, FP= False Positives, TN= True Negatives. TP/FP/TN/FN are numbers of unique mutations. The Sum column shows the number of times the method succeeded and an observation was possible and therefore reflects the robustness of the method. Sensitivity, Specificity and Accuracy defined in text.
V.5 SAV effect: SNAP predictions
Yana Bromberg, Rutgers University

SNAP1 2004-2009
Effect prediction through machine-learning
SNAP: input features: major novelty DELTA features
SNAP: input features

Prob. (TNR) >? Prob. (TLR)
SNAP1 input

SNAP

- biophysical features
- alignment profiles
- probability of residue triplets
- solvent accessibility (PROFacc)
- secondary structure (PROFsec)
- residue flexibility (PROFbval)

SNAPannotated

- SWISS-PROT annotations
- Pfam domains
- SIFT predictions

SNAP: neural network

Prob. (TNR) >? Prob. (TLR)

Score: -100 ≤ S ≤ 100

SNAP1
overall performance
## Performance comparison

Overall Two-State Accuracy (Q2)

<table>
<thead>
<tr>
<th>Method</th>
<th>LacI Repressor</th>
<th>Lysozyme</th>
<th>HIV-1 protease</th>
<th>Melanocortin-4 receptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>69.4</td>
<td>67.6</td>
<td>78.3</td>
<td>57.8</td>
</tr>
<tr>
<td>PolyPhen</td>
<td>68.7</td>
<td>57.9</td>
<td></td>
<td>51.1</td>
</tr>
<tr>
<td>SNAP</td>
<td>70.7</td>
<td>70.0</td>
<td>68.5</td>
<td>71.1</td>
</tr>
<tr>
<td>SNAP_annotated</td>
<td>72.7</td>
<td>73.2</td>
<td>72.3</td>
<td>80.0</td>
</tr>
<tr>
<td>SNPs3D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
Q2 = 100 \times \frac{\text{# correct predictions}}{\text{total # of predictions}}
\]
SNAP performance by exposure

Coverage for effect SAVs

Accuracy for effect SAVs

\[
\text{Accuracy} = 100 \times \frac{\# \text{correct predictions}}{\text{total \# of predictions}}
\]

Coverage for neutral SAVs

\[
\text{Coverage} = 100 \times \frac{\# \text{correct predictions}}{\text{total \# of observations}}
\]

SNAP reliability index

Predictions

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.8</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.49</td>
</tr>
</tbody>
</table>
SNAP RI $\sim$ severity of change

Y Bromberg & B Rost (2008) Bioinformatics 24:i207-12
## SNAP: examples

<table>
<thead>
<tr>
<th>Gene</th>
<th>nsSNP</th>
<th>Disease</th>
<th>Function</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>HXK4</td>
<td>S131P</td>
<td>Diabetes Mellitus</td>
<td>Significant increase in affinity for ATP</td>
<td>effect (2)</td>
</tr>
<tr>
<td>PAX6</td>
<td>G64V</td>
<td>Cataract</td>
<td>Reduction of DNA binding activity</td>
<td>effect (5)</td>
</tr>
<tr>
<td>MC4R</td>
<td>I301T</td>
<td>Obesity</td>
<td>Severe change of basal activity &amp; EC50</td>
<td>effect (3)</td>
</tr>
<tr>
<td>HXK4</td>
<td>M107T</td>
<td>Polymorphism</td>
<td>Not conclusive</td>
<td>neutral (0)</td>
</tr>
<tr>
<td>CFTR</td>
<td>P1013</td>
<td>Cystic Fibrosis</td>
<td>Not conclusive</td>
<td>effect (5)</td>
</tr>
<tr>
<td>NKX2</td>
<td>A127G</td>
<td>Secundum atrial septal defect</td>
<td>Not conclusive</td>
<td>neutral (6)</td>
</tr>
<tr>
<td>HBB</td>
<td>R104T</td>
<td>Polymorphism</td>
<td>Not conclusive</td>
<td>effect (3)</td>
</tr>
<tr>
<td>P53</td>
<td>R337H</td>
<td>Adrenocortical carcinoma</td>
<td>Does not affect transactivation</td>
<td>effect (3)</td>
</tr>
</tbody>
</table>
Annotations help

- Annotation
- Alignment
- All

Percentage of SNPs predicted correctly

Annotations help, but not often

Crucial sites identified in insulin

http://www.rostlab.org/servers/SNAP/

Y Bromberg & B Rost 2007 NAR 35:3823-35

Y Bromberg G Yachdav & B Rost 2008 Bioinformatics 15:2397-8

model: SWISS-MODEL
SNAP: glucokinase


Y Bromberg & B Rost (2008) Bioinformatics 24:i207-12
SNAP 1 performance in comparison
SNAP performs well

80,000 mutants w/known effects on function

Predict Features

Q2

Random
SIFT
PolyPhen
SNAP

\[ Q2 = 100 \times \frac{\text{# correct predictions}}{\text{total # of predictions}} \]

© Yana Bromberg, 2010 Columbia University

Y Bromberg & B Rost (2008) Bioinformatics 24:i207-12
Y Bromberg, G Yachdav & B Rost (2008) Bioinformatics 15:2397-8
SNAP clearly best for subtle classes

<table>
<thead>
<tr>
<th>PMD/EC data</th>
<th>Unknown</th>
<th>Accuracy effect</th>
<th>Coverage effect</th>
<th>Accuracy neutral</th>
<th>Coverage neutral</th>
<th>Overall two-state accuracy (Q2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>2374 (3%)</td>
<td>79.8±0.6</td>
<td>63.4±1.2</td>
<td>70.1±2.7</td>
<td>84.3±1.2</td>
<td>74.0±1.4</td>
</tr>
<tr>
<td>PolyPhen</td>
<td>1647 (2%)</td>
<td>79.1±0.7</td>
<td>66.9±1.4</td>
<td>71.8±2.7</td>
<td>82.7±1.1</td>
<td>74.9±1.3</td>
</tr>
<tr>
<td>SNAP</td>
<td>0</td>
<td>76.3±0.8</td>
<td>83.3±1.0</td>
<td>82.0±2.4</td>
<td>74.7±2.2</td>
<td>78.9±1.3</td>
</tr>
</tbody>
</table>

Accuracy = $100 \times \frac{\text{# correct predictions}}{\text{total # of predictions}}$

Coverage = $100 \times \frac{\text{# correct predictions}}{\text{total # of observations}}$

Y Bromberg & B Rost (2008) Bioinformatics 24:i207-12
Y Bromberg, G Yachdav & B Rost (2008) Bioinformatics 15:2397-8

© Yana Bromberg, 2010 Columbia University
SNAP performs well, but…

\[
Q2 = 100 \times \frac{\text{# correct predictions}}{\text{total # of predictions}}
\]

Y Bromberg & B Rost (2008) Bioinformatics 24:i207-12
Y Bromberg, G Yachdav & B Rost (2008) Bioinformatics 15:2397-8
SNAP much better for tough cases

\[ Q2 = 100 \times \frac{\# \text{correct predictions}}{\text{total # of predictions}} \]

- Y Bromberg & B Rost (2008) Bioinformatics 24:i207-12
SNAP²

Predicting functional effects of sequence variants

2010-2015

M Hecht, Y Bromberg & B Rost (2013) JMB
SNAP2 Training Data

- Training data consisting of ~100,000 variants
  - Protein Mutant Database (PMD, experimental annotation)
  - Enzyme Classification (EC, putative neutrals from known enzymes)
  - Disease (DISEASE, disease-related variants)

<table>
<thead>
<tr>
<th>Source</th>
<th>Effect</th>
<th>Neutral</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMD</td>
<td>38k</td>
<td>13k</td>
<td>51k</td>
</tr>
<tr>
<td>EC</td>
<td>-</td>
<td>27k</td>
<td>27k</td>
</tr>
<tr>
<td>DISEASE</td>
<td>23k</td>
<td>-</td>
<td>23k</td>
</tr>
<tr>
<td>Total</td>
<td>61k</td>
<td>40k</td>
<td>101k</td>
</tr>
</tbody>
</table>

- Independent validation set
  - *E. coli* LacI repressor (4,041 variants)

  M Hecht, Y Bromberg & B Rost (2015) BMC Genomics 16:S1
Prediction Features

Secondary Structure

Annotation

Contacts & co-evolution

Physicochemical features

Evolutionary information

© Burkhard Rost
SNAP1 input

☐ SNAP
  ● biophysical features
  ● alignment profiles
  ● solvent accessibility (Reprof)
  ● secondary structure (Reprof)
  ● residue flexibility (PROFbval)
  ● AAindex + other global features
  ● predicted binding sites (InteractionSites, DIS)
  ● disordered regions
  ● contact potentials
  ● correlated mutations
  ● low-complexity regions

Special versions
  • no alignment
  • no disease data

☐ SNAPannotated
  ● SWISS-PROT annotations
  ● Pfam domains
  ● PROSITE
  ● SIFT predictions

SNAP2 improves throughout

Accuracy = $100 \times \frac{\text{# correct predictions}}{\text{total # of predictions}}$

Coverage = $100 \times \frac{\text{# correct predictions}}{\text{total # of observations}}$

SNAP2 better than naïve combinations

M Hecht, Y Bromberg & B Rost (2015) BMC Genomics 16:S1:
doi:10.1186/1471-2164-16-S8-S1
SNAP2 best for tough human variants

• High agreement between methods: 61%-77%

• Some predictions are more difficult than others

• Classification (84k):
  • Easy (54k)
  • Unsolvable (6k)
  • Difficult (24k)

\[ Q2 = 100 \times \frac{\# \text{ correct predictions}}{\text{total } \# \text{ of predictions}} \]
Accuracy and Reliability

Prediction of orphan variants

- 69% prediction accuracy on training data
- Alignment information is important

Prediction of orphan variants

- No real orphans in our data
- But: 8 Proteins (248 variants) with less than 5 hits
- More data needed for significant results

<table>
<thead>
<tr>
<th>SNAP2noali</th>
<th>SNAP2</th>
<th>PolyPhen-2</th>
<th>SIFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>62%</td>
<td>61%</td>
<td>60% (104)</td>
<td>N/A</td>
</tr>
</tbody>
</table>

preliminary Lecture plan (PP2 function)

01: 2018/10/16: No lecture (makeup examen; PP last year)
02: 2018/10/18: No lecture (makeup)
03: 2018/10/23: Welcome: who we are
04: 2018/10/25: Intro function 1: concept of protein function
05: 2018/10/30: No lecture
06: 2018/11/01: No lecture (holiday, All Saints)
08: 2018/11/08: Localization 1 (chalk talk)
09: 2018/11/13: No lecture (SVV)
10: 2018/11/15: Localization 2 (homology, motifs)
11: 2018/11/20: Localization 3 (motifs, machine learning)
12: 2018/11/22: No lecture (Thanksgiving)
14: 2018/11/29: Localization 5 (machine learning 2)
15: 2018/12/04: Localization 6
16: 2018/12/06: No lecture (Dies Academicus)
17: 2018/12/11: PPI 1 - sites / pairing (chalk)
18: 2018/12/13: PPI 2 - sites / PPI pairing (chalk)
19: 2018/12/18: PPI 3 - sites / DNA / RNA (Jia Jun Qiu)
20: 2018/12/20: No lecture
21-24: no lectures - winter break (2018/12/24 - 2019/01/06)
25: 2019/01/08: No lecture
28: 2019/01/10: PPI 4 - sites: DNA / RNA (Jia Jun Qiu) + PPI pairing 1
29: 2019/01/15: PPI 5 - PPI pairing 2
30: 2019/01/17: SNV effect 1 (chalk talk)
31: 2019/01/22: SNV effect 2
32: 2019/01/24: SNV effect 3
33: 2019/01/29: WRAP up 1
34: 2019/01/31: WRAP up 2
35: 2019/02/05: ?
36: 2019/02/07: Examen (10:00-13:00, lecture room LMU physics)