V. Predict effect of sequence variation

cb2_sav1

lecture: Computational Biology 2 - Protein function (for Informatics) - TUM summer semester
Announcements

Videos: YouTube / www.rostlab.org

THANKS:

Special lectures:
- 11/21 THU TBC: Konstantin Weissenow: deep learning binding sites
- 12/12 THU TBC: Maria Littmann & Jia Jun Qiu: Protein-DNA-RNA-ligands binding
- 01/21 TUE Dmitirii Nechaev: Deep learning SAV predictions

No lecture:
- 10/31 THU All Saints
- 11/12 TUE SVV (student rep)
- 11/28 THU Thanksgiving
- 12/05 THU TUM Dies Academicus
- 12/19-01/07 - no lecture Xmas+

LAST lecture: Jan 21 (followed by 2 wrap-up sessions)

Examen: Jan 30 11:00-13:00 Room LMU Physics HS019
- Makeup: NONE (emergency: Apr 21 & Apr 23, 2020 lecture time)
V. Predict effect of mutations
V.1 SNP|SNV/SAV 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**
  - 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
Types of sequence variation - Protein level

- **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)
Sequence variation - nomenclature

- SNV: Single Nucleotide Variant
  previously: SNP
- “Mutation”
  - A change, possibly disease-causing, not well defined
- SAV: Single amino acid variant (SAV)
- 19 non-native
  - Substituting the native amino acid by all others
  - In contrast to SNV-possible
### SNV-possible

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

---

<table>
<thead>
<tr>
<th>1st base</th>
<th>2nd base</th>
<th>3rd base</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>U</td>
<td>U</td>
</tr>
<tr>
<td>U</td>
<td>C</td>
<td>U</td>
</tr>
<tr>
<td>U</td>
<td>A</td>
<td>U</td>
</tr>
<tr>
<td>U</td>
<td>G</td>
<td>U</td>
</tr>
<tr>
<td>U</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>C</td>
<td>U</td>
<td>C</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>A</td>
<td>U</td>
<td>U</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>G</td>
<td>U</td>
<td>U</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>G</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
</tbody>
</table>

---

https://en.wikipedia.org/wiki/Genetic_code
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!
V.2 SAV effect:
First 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

Test Set: 3 proteins, ~6500 mutants
PolyPhen: Polymorphism Phenotyping
PolyPhen

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

Train/Test Set: 3768 HGMD variants and 2309 cross-species
variants
P Yue, Z Li & J Moult (2005)= Loss of protein structure stability as a
major causative factor in monogenic disease. JMB 353: 459-63.
nsSNP effects: some in silico methods

- **SIFT**
  - PC Ng & S Henikoff (2003) NAR 31:3812-14
  - Sequence:
    - VHLTPEEKSA
    - VTALWGBKVN
    - DEVGEALGR
    - LLVYYPWTQR
    - FFESFGDLST
    - PDAVMGNPKV
    - KAHGKKVLGA
  - Mutant: E6V

- **PolyPhen**
  - Sequence:
    - VHLTPEEKSA
    - VTALWGBKVN
    - DEVGEALGR
    - LLVYYPWTQR
    - FFESFGDLST
    - PDAVMGNPKV
    - KAHGKKVLGA
  - Mutant: E6V

- **SNPs3D**
  - P Yue, Z Li & J Moult (2005) JMB 353:459-63
  - Sequence:
    - VHLTPEEKSA

© Burkhard Rost
V.3 SAV effect: Data
Misfunction/neutral
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?
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 YQQLFELMNKVGAFSRLRLKEHTHTFVNKGGRTGALDFRFTGAPFNGLKAFFTTSQSL 120
       YY LF LM KVGA +LRLKEHTHTFVN+GGR G LDFRF TGAPFNLKAFFTTSQ

Sbjct: 61 YQNLFNLMEKVGAQNLRLKEHTHTFVNQGGRIGEFLDFRFTGAPFNLKAFFTTSQLDT 120

Neutral: 26,840


## Availability of data for nsSNP limiting

<table>
<thead>
<tr>
<th>Category</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>26,840</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>40,641</td>
<td>41,174</td>
</tr>
</tbody>
</table>

**81,815 variants**

**6,821 proteins**

---

V.4 SAV effect: 2nd 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*)

<Image of Yana Bromberg>

SNAP²
Predicting functional effects of sequence variants

Enter protein sequence in FASTA format

Optional: Email address

RUN PREDICTION  EXAMPLE1  EXAMPLE2  CLEAR

Copyright © 2014 ROSTLAB all rights reserved.
Some other methods in quick
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>Sequence</th>
<th>Subfamily</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGFR_HUMAN</td>
<td>GLKELPMRNLOGEILHGAVRFSNN</td>
<td>subfamily 1</td>
</tr>
<tr>
<td>Q8M18_PIG</td>
<td>GLRELPMRNLOGEILHGAVRFSNN</td>
<td></td>
</tr>
<tr>
<td>EGFR_CHICK</td>
<td>GLRELPMKRLGEILNGYKISNN</td>
<td></td>
</tr>
<tr>
<td>ERBB4_HUMAN</td>
<td>GKEQLGKNLTEILNGGYYDQN</td>
<td></td>
</tr>
<tr>
<td>INSR_MOUSE</td>
<td>HKEGLGLLNMDTSVRIEKNN</td>
<td>subfamily 2</td>
</tr>
<tr>
<td>ILPR_BRALA</td>
<td>DMEKGLYSLQNTSVRIEKNN</td>
<td></td>
</tr>
<tr>
<td>IGFR1_XENLA</td>
<td>DMEKIGLYNRNITNGAVRIKNN</td>
<td></td>
</tr>
</tbody>
</table>

**Functional Impact Score**

\[ \text{Functional Impact Score} = \text{conservation score} + \text{specificity score} \]

**Validation**

**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
Evolutionary Couplings

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

single column frequencies: $f_i(A_i) \quad f_j(A_j)$

column pair frequencies: $f_{ij}(A_i, A_j)$

\[ f_{ij}(A_i, A_j) = f_i(A_i)f_j(A_j) \]
Evolutionary Couplings predict effects of variants

T Hopf, C Sander, D Marks (2016) submitted
How to compare methods?
V.5 SAV effect: SNAP predictions
SNAP
Effect prediction through machine-learning
SNAP: input features: major novelty DELTA features
SNAP: input features

\[ \text{Prob. (TNR)} > ? \text{ Prob. (TLR)} \]
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

SNAP 1
overall performance
## Performance comparison

### Overall Two-State Accuracy (Q2)

<table>
<thead>
<tr>
<th>Method</th>
<th>Lacl 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><strong>78.3</strong></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>SNAPannotated</td>
<td><strong>72.7</strong></td>
<td><strong>73.2</strong></td>
<td><strong>72.3</strong></td>
<td><strong>80.0</strong></td>
</tr>
<tr>
<td>SNPs3D</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>62.2</td>
</tr>
</tbody>
</table>

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

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

Coverage for effect SAVs

Accuracy 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>0.8</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.49</td>
<td></td>
</tr>
</tbody>
</table>
SNAP reliability index

Cumulative accuracy vs. Cumulative percentage predicted with $RI \geq n$

- Red: non-neutral
- Green: neutral

SNAP RI ~ severity of change

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

Percentage of SNPs predicted correctly:

- Annotation: 80%
- Alignment: 70%
- All: 60%

Annotations help, but not often

SNAP1 performance in comparison
SNAP performs well, but…

\[
Q_2 = 100 \times \frac{\# \text{correct predictions}}{\text{total # of predictions}}
\]

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

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

**SNAPannotated**
- SWISS-PROT annotations
- Pfam domains
- PROSITE
- SIFT predictions

Special versions
- no alignment
- no disease data
SNAP2 improves throughout

Accuracy = 100 * \( \frac{\text{# correct predictions}}{\text{total # of predictions}} \)

Coverage = 100 * \( \frac{\text{# correct predictions}}{\text{total # of observations}} \)

SNAP2 better than naïve combinations

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

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)

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

Accuracy correlates with reliability (prediction strength)
Lecture plan (CB2 function)

01: 2019/10/15: No lecture (makeup examen: PP last year)
02: 2019/10/17: No lecture (makeup)
03: 2019/10/22: No lecture
04: 2019/10/24: Welcome: who we are
05: 2019/10/29: Intro function 1: concept of protein function
06: 2019/10/31: No lecture (holiday, All Saints)
07: 2019/11/05: Localization 1 (chalk talk)
08: 2019/11/07: Localization 2 (homology, motifs)
09: 2019/11/12: No lecture (SVV)
10: 2019/11/14: Localization 3 (machine learning)
11: 2019/11/19: Localization 4 (machine learning)
13: 2019/11/26: Localization 5 (machine learning)
14: 2019/11/28: No lecture (Thanksgiving)
15: 2019/12/03: PPI 1 - sites (chalk)
16: 2019/12/05: No lecture (Dies Academicus)
17: 2019/12/10: No lecture
18: 2019/12/12: PPI 2 - sites / DNA / RNA (Jia Jun Qiu) / small molecules (Maria Littmann)
19: 2019/12/17: PPI 3 - sites
20: 2019/12/19: No lecture
22-24: no lectures - winter break (2019/12/24 - 2020/01/06)
25: 2020/01/07: No lecture
28: 2020/01/09: PPI 4 - pairs
29: 2020/01/14: SAV effect 1
30: 2020/01/16: SAV effect 2
31: 2020/01/21: SAV effect 3 - Dmitrii Nechaev Deep learning SAV prediction
32: 2020/01/23: WRAP up 1
33: 2020/01/28: WRAP up 2
34: 2020/01/30: Examen (11:00-13:00, lecture room HS019 LMU physics)
35: 2020/02/04: TBA
36: 2020/02/06: TBA

© Burkhard Rost