Machine Learning Methods in Computational Biology

Using RVM probabilistic classification to list of diseases

Nikolas Begetis

Single nucleotide polymorphism mutations

- Motivation:
  - Find related pathogenic SNPs, categorized in distinct disease families and classify them to notice if there is a connection among them (test in pairs one disease to all others).

- Data:
  - Identified SNPs taken from the dbSNP data bank of NCBI
  - http://www.ncbi.nlm.nih.gov/snp

Data

List of Diseases

- Cancer types
- Cutaneous conditions
- Endocrine diseases
- Eye diseases and disorders
- Intestinal diseases
- Infectious diseases
- Communication disorders
- Genetic disorders
- Neurological disorders
- Voice disorders
- Vulvovaginal disorders
- Mental illness

Verified SNP connections with diseases

Using GeneCards and NCBI dbSNP

- Cancer types (424/217)
- Cutaneous conditions (83/42)
- Endocrine diseases (70/35)
- Eye diseases and disorders (399/56)
- Intestinal diseases (39/20)
- Infectious diseases (94/49)
- Communication disorders (12/6)
- Genetic disorders (6/3)
- Neurological disorders (278/139)
- Voice disorders (10/5)
- Vulvovaginal disorders (2/1)
- Mental illness (550/275)

SNP Information Gathering

- **Brief**
  - Brief
  - FASTA
  - FlatFile
  - Summary
  - Chromosome Report
- **Detailed**
  - ASN.1
  - XML

Use of DOM for XML parsing and BioJava for Feature Extraction

- XML Example Screenshot

```
<Ss ssId="537712977" handle="NCBI-CURATED-RECORDS" batchId="1057285" locSnpId="23272" subSnpClass="snp" orient="forward" strand="top"
    <Sequence>
        <Seq5>GAAATGTGCTACACTGGACACCAGAGCCCTGCTTAAATTGGCTTATTCAG</Seq5>
        <Observed>A/C</Observed>
        <Seq3>TGCCTGTGGTCTGTGCCACCTGCACACAGAAATTTATGGCACCCAAGGAA</Seq3>
    </Sequence>
</Ss>
<Assembly dbSnpBuild="137" genomeBuild="37.4" groupLabel="GRCh37.p9" current="true" reference="true">
    <Component componentType="contig" accession="NT_030059.13" chromosome="10" start="49195536" end="128616068" orientation="fwd" gi="
        <MapLoc asnFrom="39483536" asnTo="39483536" locType="exact" alnQuality="1" orient="forward" physMapInt="88679072" leftContigNe
            <FxnSet geneId="657" symbol="BMPR1A" mrnaAcc="NM_004329" mrnaVer="2" protAcc="NP_004320" protVer="2" fxnClass="missense" r
            <FxnSet geneId="657" symbol="BMPR1A" mrnaAcc="NM_004329" mrnaVer="2" protAcc="NP_004320" protVer="2" fxnClass="reference"
        </MapLoc>
    </Component>
    <SnpStat mapWeight="unique-in-contig" chromCount="1" placedContigCount="1" unplacedContigCount="0" seqlocCount="1" hapCount="0"/>
</Assembly>
```

Features

- 148 features from the XML file
- + 15-25 features given from BioJava information elaboration

Row data → Features

- seq5 "GAAATGTGCTACACTGGACACCAGAGCCCTGCTTAAATTGGCTTATTCAG"
  - 0...100% για GC content κλπ
- observed "A/C"
  - {1...16} σύμφωνα με το πρότυπο IUPAC (http://www.bioinformatics.org/sms2/iupac.html)
- seq3 "TGCCTGTGGTCTGTGCCACCTGCACACAGAAATTTATGGCACCCAAGGAA“
  - 0...100% για GC content κλπ
- ssId 506984377
  - {1...n} σε ένα hash map με id, εξετάζοντας αν τα ids είναι ίδια σε περισσότερες από μία ομάδες ασθενειών
- batchId 1056763
  - {1...n} σε ένα hash map με id, εξετάζοντας αν τα ids είναι ίδια σε περισσότερες από μία ομάδες ασθενειών
- orient forward
  - {0/1} 0 για backward, 1 για forward
- strand top
  - {0/1} 0 για bottom και 1 για top
- molType genomic
  - {1...n} όπου n όλες οι διαφορετικές κατηγορίες
- validated by-submitter
  - {0/1} 0 χωρίς validation 1 με validation

More Features

- Features selection:
  - Effects on physiochemical alternation(hydrophobic, polar, charged, glycine)
  - Sequence conservation score at the mutated position
  - Molecular mass shift on mutation
  - Hydrophobicity difference
  - Secondary structure
  - Codon from which they came
  - Aminoacid composition: # aminoacids altered by snp
  - GC content
  - Ratios: A/T, AT/GC
  - Motif search Position in gene
  - Pyrimidine/Purine
  - Any information we can relate to snp or its sequence (bases, codons, aminoacids, alleles)

References and Links

- http://www.snpedia.com/index.php/SNPedia
- http://www.ncbi.nlm.nih.gov/snp
- ftp://ftp.ncbi.nih.gov/pub/factsheets/Factsheet_SNP.pdf
  - Manual for the dbSNP
- http://en.wikipedia.org/wiki/Lists_of_diseases
- http://gene4.weizmann.ac.il/
  - Validations of existance in certain disease family
- http://www.bioinformatics.org/sms2/iupac.html
  - IUPAC fasta format
- http://www.ncbi.nlm.nih.gov/books/NBK44476/
  - ASN.1 and xml format usage