Machine approaches to recognition of trypanosomal variant surface glycoprotein sequences

Jonathan M. Wilkes

Trypanosomes – Variant Surface Glycoprotein properties

• Cell membrane covered in a dense glycocalyx – excludes HMwt molecules (Abs).
• Consists of VSGs: GPI anchored homodimers. Highly immunogenic.
• Stochastic switching of expressed VSG enables evasion of host immune response.
• Novel sequences produced by recombination from archive containing >2000 seqs.
• Only 5% of archive codes for full length, functional protein. 95% genes, fragments.
• Sequences within the archive display <20% sequence identity.
• Similarity and profile methods are ineffective for detection of VSGs in genomic seqs
• Next Generation Sequencing producing prodigious amount of sequence
• VSGs of genome strain (TREU 927) extensively hand annotated; N- and C-terminal domains characterised (VSGdb.net) – source of training & test sequences.
SVM -

- supervised machine learning method for the classification of data
- each datum of the training set is expressed as a category label, and a vector of numerical values (features) derived from the properties of the objects of interest (attributes)
- algorithm projects the vectors into higher dimensional spaces—determines an optimal (hyper)-plane separating the categories, highest weighting given to support vectors. Non-linear distributions of categories can be transformed by the use of kernel functions
- capable of determining the category of a test datum based on the attribute vector derived from it
- I have developed a suite of perl modules (SVMseq) to automate the task of feature selection from a set of biological sequences:
  1. utilises Algorithm::SVM module, radial basis function as kernel
  2. permits rapid prototyping and combination of features.
  3. Automates scaling of vector elements in range -1 to +1
  4. combines cross-validation and directed grid-search to optimize model parameters ($\gamma$, $C$).
  5. ROC curve analysis to determine discriminatory power (insensitive to large excess of negative over positive samples in training/test set).

See Poster

Virtual potential transformation can detect VSG sequences in both senses

Vectorisation by compound virtual potential transform [CC+CG+GC+GG]

- ROC analysis
- SVMscore normalised cumulative frequency
- AUC = 0.967
- AUC = 0.962
Detection of C- and N-terminal domains in established VSG arrays

Overlapping VSG containing sequences merged into ‘VSG arrays’, which were subsequently processed for detection of C- and N-terminal domains using a variety of vectorization techniques.

Discrimination of N- and C-terminal domains by the SVMcodonScatter vectorization for the 20 amino acids

- Most transforms have higher discrimination for N-terminal domains
- Glutamine (Q) unexpectedly good discriminator
- Cysteine (C) is a poor discriminator despite its structural significance

Discrimination of N- and C-terminal domains with other vectorization methods

- Both N-terminal (red, 600bp sequences) and C-terminal (green, 300bp sequences) show highest discrimination with SVMspectrumN (N=3) (ie frequency distribution of all 64 triplets in the sequence)

Mapping of SVM positive sequences to genomic sequence

Analysis of a test data set (chromosome 9 of TREU927), showing the distribution of annotated VSGs (inner lanes; true positives), putative VSG arrays (middle lanes, yellow blocks derived from forward & reverse VPs) and predictions of C-(green) and N-(red) terminal domain sequences (outer lanes, SVMspectrum_N, N=3 of the putative VSG array sequences only).

Novel annotation of VSGs in T. brucei gambiense

Contig 1576d_6_q1k from T.b.gambiense DAL 927 sub-telomeric ‘bin’ sequence set. Mapping of predicted N- and C-terminal sequences utilizing the models trained on T.b.brucel TREU927. Note detection of a fragment of a VSG array, visualized with ARTEMIS