HapRFN: a deep learning method for identifying short IBD segments

Gundula Povysil1, Djork-Arné Clevert2, Sepp Hochreiter1
1 Institute of Bioinformatics, Johannes Kepler University Linz, Austria, 2 Bayer Pharma AG, Berlin, Germany

Abstract

A segment of DNA is called identical by descent (IBD) in two or more individuals if it is identical because it was inherited from a common ancestor. IBD segments can be used to uncover hidden familial relationships, detect the population of origin of an individual or find interbreedings between humans and ancient hominids like the Neanderthal. IBD segments can also be used to find the cause of diseases via a technique called IBD mapping.

For the above applications HapFABIA was shown to be superior to other IBD detection methods by detecting short IBD segments systems with very rare DNA variants via biclustering. However, HapFABIA still has several problems: (1) To decide whether individuals possess an IBD segment is often difficult because of the soft bicluster membership supplied by HapFABIA. (2) HapFABIA can only extract 10-30 IBD segments at once and therefore needs to perform multiple iterations. (3) The IBD segments identified in different iterations may not be de-correlated, thus they may be redundant and overlapping or even split into smaller segments. (4) Processing very large data sets is time intensive.

We recently introduced "Rectified Factor Networks" (RFNs) as an unsupervised deep learning approach. Each code unit of the RFN represents a bicluster and therefore an IBD segment, where samples for which the code unit is active share the bicluster (IBD segment) and features (DNA variants) that have activating weights to the code unit tag the IBD segment. HapRFN overcomes the problems of HapFABIA. (1) RFNs provide sparser codes via their rectified linear units that immediately supply bicluster memberships as factors being different from zero. (2) RFNs can learn thousands of factors and therefore many IBD segments more precisely. With HapRFN we are able to accurately detect familial relationships, populations of origin, or interbreeding with ancient genomes in data sets with thousands of individuals. Furthermore, finding disease associations via IBD mapping becomes more reliable which might be the key to uncover unknown hereditary causes of multifactorial diseases.

Identity by Descent (IBD)

IBD (identity by descent): identical DNA sequence in 2 or more individuals because it was inherited from a common ancestor

SNVs (single nucleotide variants):
small differences between DNA sequence of sample and reference, can be used to distinguish individuals, shared if sequence is IBD

Rectified Factor Networks:
• \( v = \mathbf{W} h + \epsilon \) - visible units/observations
• \( h \sim \mathcal{N}(0, I) \) - hidden units
• \( W \in \mathbb{R}^{N \times K} \) - weight matrix
• \( \epsilon \sim \mathcal{N}(0, \Psi) \) - additive noise
• rectified posterior mean \( \tilde{h} \)
• non-negative and exact zero code units
• dropout \( \rightarrow \) sparse code units
• generalized alternating minimization algorithm
• based on posterior regularization method

Biclustering – IBD Segments

bicluster: pair of sample set and feature set for which samples are similar to each other on the features and vice versa

IBD segment = bicluster to which
• SNVs (features) belong if they tag the IBD segment (tagSNVs)
• individuals (samples) belong if they share the IBD segment

HapFABIA

• biclustering of SNV data to find short IBD segments shared by multiple individuals
• based on low-frequency and rare SNVs (minor allele frequency < 5%)

Problems:
(1) soft bicluster membership: deciding whether individuals possess IBD segment is difficult
(2) multiple iterations: HapFABIA can only extract 10-30 IBD segments at once, IBD segments identified in different iterations may not be de-correlated, may be redundant and overlapping or even split into smaller segments
(3) time intensive

HapRFN

• each code unit of RFN represents bicluster/IBD segment
• individuals for which code unit is active share IBD segment
• features (SNVs) that have activating weights to code unit tag IBD segment

HapRFN overcomes the problems of HapFABIA:
(1) RFNs provide sparser codes via their rectified linear units that immediately supply bicluster memberships as factors being different from zero.
(2) RFNs can learn thousands of factors and therefore many IBD segments simultaneously.
(3) RFNs are much faster due to techniques from deep learning like efficient matrix multiplications and implementations of networks on graphical processing units (GPUs).

Conclusion

• rectification \( \rightarrow \) easier interpretation of bicluster membership
• simultaneous detection of thousands of IBD segments
• better separation of IBD segments \( \rightarrow \) less intensive post-processing
• main results unchanged but more reliable


Acknowledgements: We thank the NVIDIA corporation for the GPU donation that facilitated our work.