IgFamily v0.12.3

Flinders Proteomics Facility

Contents

IgFamily

Mass spectrometry has the potential to investigate serum immunoglobulin repertoires as an accurate and high-throughput method. Particularly applicable to patients with immunological maladies, immunoglobulins are able to be selectively purified using an antigen or affinity-specific molecule. Typically a sample of immunoglobulin proteins is digested with an enzyme to produce peptides that are suitable for mass spectrometry analysis. Such a sample is able to generate many thousands of spectra that are representative of the peptides present. The entirety of spectra allows insight into the origin proteins of these peptides and the abundance of proteins in the sample. Conventional analysis of observed peptides involves determining those peptides that best represent the possible proteins, often from a database of established proteins, through a measure of spectral quality, spectral count, unique and supporting peptides, and overall protein sequence coverage.

Although this has proven successful for general protein studies there are complications implicit for immunoproteomics. In particular, the comparative database of immunoglobulins is derived from an ancestral repetition of successful germline alleles referred to as immunoglobulin gene families. Of these the variable region alleles are responsible for creation of the ~100 amino acid N-terminal and are of importance in diversification of target affinity. There are spatially 73 heavy chain variable region alleles, each with many known polymorphic allele variants and potentially many more not yet described. The variable allele has regions of relatively strong germline conservation separated by phylogenetically significant regions of diversity. In addition, owing to a RNA polymerase of relatively low mismatch fidelity, there are hypervariable regions with a somatically increased nucleotide mutation rate. The generally short-length peptides generated through a mass spectrometry method often results in peptides corresponding to regions of database proteins that are on one case strongly conserved but on the other prone to excessive germline and mutagenic variation. These confounding factors raise concerns of the validity of current mass spectrometry derived immunoproteomic approaches.

The IgFamily program was developed to automate the data management and inference workflow of large bodies of mass spectrometry generated data. The initial concept was to emulate the conventional approach of focusing inferential effort on unique peptides representative of the associated proteins. As the program was developed interest grew in the role of supporting peptides and what could be revealed about the holistic complexity of the data. The product of this was a model that could determine the likelihood of each peptide originating from any of the database proteins, while also taking into consideration the overall evidence contained in the data. The role of the data and a discussion of the model is provided in the Story chapter. The functionality of the IgFamily program is described in the Function chapter.

An example is considered in the Consideration chapter. The majority of observed peptides are difficult to assign by having a broad possibility of originating proteins and assignment of these peptides is aided through the resolving power of distinct peptides. Interestingly, as a product of the model the clonal divergence becomes apparent. A key observation follows that the amino acid substitution is more favourable to those changes that are similar to other gene family germlines. Two gene families in the example have unique peptides and a continuum of germline divergent peptides between them. With a conservative rate of amino acid substitution peptides of this grouping are assigned to either of the gene families. With a more vigorous substitution rate, most peptides and the unique peptide for one of the gene families may be explained as having an origin from the other.

Story

- Data: what is it?

Protein mass spectrometry is capable of producing thousands of spectra from a sample of peptides. Although the spectra are representative of the peptides present, correctly assigning a peptide to a spectrum is a complex task. The two primary methods of peptide assignment are through database matching and by de novo identification. Database matching relies on an established database for which an in silico enzymatic digest of database proteins is able to produce peptide candidates that are compared to the fragmentation spectra. De novo identification assumes little about the sample and attempts to assign a peptide on the basis of probable amino acid fit. The PEAKS software suite uses a combination of de novo and database matching to determine protein likelihood and is able to produce data for each of these methods independently (Zhang, et al., 2012). In addition, the open-source NOVOR program can assign peptides to spectra through a de novo method. The IgFamily program is able to analyse both PEAKS de novo and database assigned peptides as well as peptides assigned by the NOVOR de novo program.

- Data: where does it come from?

The primary role of the IgFamily program is to determine the proteins present in a sample. In a sense, spectral assigned peptides provide a snapshot of their origin proteins. Although in general a greater quantity of supporting peptides bolsters the likelihood of the existence of a protein, not all peptides have equal discerning power. There are two primary considerations:

- How much of the peptide is present? - What does this peptide belong to?

The first question concerns a complex array of processes. Elution from the chromatographer may result in an otherwise abundant peptide to produce few spectra. Coelution may shroud the quantity of an individual peptide and result in less dedicated mass spectrometer cycle time. Fragmentation efficiency can give a misrepresentative idea of the proportion of peptides present or result in some existing peptides to not be observed altogether. These are a small example of variations that understate the importance of peptide abundance in determining sample proteins. However, the IgFamily program considers only spectral count as a metric of abundance and encourages the user to keep these complications in mind.

The second question involves the certainty that a peptide can be assigned to a protein given the available evidence. Considering a potential peptide match requires defining the likelihood of assignment. The *conditional probability* of a peptide $\boldsymbol{p}$ originating from a protein $\boldsymbol{g}$ is defined as -

$$\boldsymbol{p(g \mid p)}$$

For example, $\boldsymbol{p(IGHV3\text{-}23 \mid TAVVYCAR) = 0.3}$ states that the probability that the peptide TAVVYCAR originating from IGHV3-23 is 0.3 or 30%. There exists a distinct conditional probability for each of all peptides $\boldsymbol{p} \in \boldsymbol{P}$ to originate from each of all proteins $\boldsymbol{g} \in \boldsymbol{G}$.

Sequence similarity provides a measure of assignment to a protein and in general the determination of the conditional probability $\boldsymbol{p}(\boldsymbol{g} \mid \boldsymbol{p})$ is a routine task. There are often segments of a peptide that are easily distinguishable to a protein even when compared to a large database. Isoform variants have the potential to confound the assignment, although there are likely to be additional supporting peptides to discern between them. However, the proteins of interest for the IgFamily program are immunoglobulins, and the nature of immunoglobulin diversification creates an almost continuum of possible peptide associations (Figure 1). The variable region of an immunoglobulin is partitioned into six defined regions: three framework regions (FR1, FR2, and FR3) separated by three complementarity-determining regions (CDR1, CDR2, and CDR3). The phylogenetic germline of these regions are known to be more conserved in the FRs, while the CDRs show greater ancestral divergence. In addition, mutation of mature B-cell germline occurs at an accelerated rate in the CDRs, owing to a RNA polymerase of relatively low mismatch fidelity. As a consequence the spatial location of a peptide has a direct impact on the ability to resolve its origin protein.

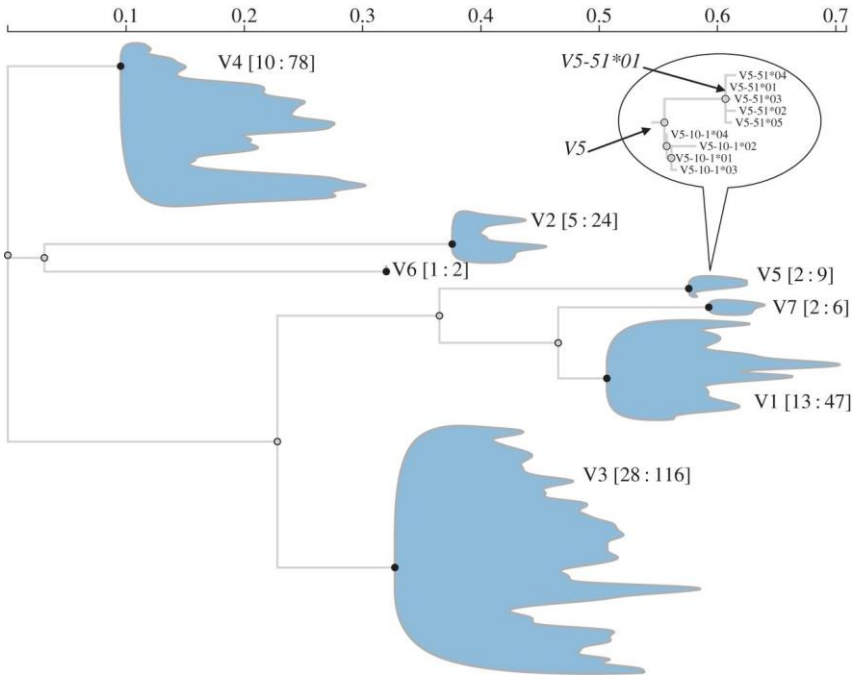


Figure 1 - The immunoglobulin gene family ancestry (Frost, et al., 2015). Phylogeny was determined through a maximum-likelihood method

- Data: when is it good?

An initial step for the analysis of assigned peptides is to filter out data that is not useful or as a worst case spuriously misleading. Consider the PEAKS de novo assigned peptides in Table 1. For each spectrum the PEAKS de novo algorithm assigns a peptide along with an associated local confidence scores for proposed residues. Often the terminal ends of a peptide produce poor fragmentation, evident in Table 1 by local confidence values below 60 for the two leading N-terminal residues. Furthermore, there are assigned peptides with poor C-terminal residues. To extract quality data two methods are proposed - Selecting peptide assignments on the basis of average local confidence (ALC) and filtering individual residues by local confidence of a peptide.

Table 1 - Peptide filtering through a de novo local confidence rolling average method report for the file WM16_B02+B02a_HAGG_ISOLATED_RF_MJ2. Here the threshold average value has been set to 85%.

p_mz	p_rt	p_withmod	p_filtered	p_denovo_peptide_local_confidence
575.819	38.65	SALVTVSSASTK	LVTVSSASTK	S[45]A[54]L[98]V[99]T[99]V[99]S[98]S[96]A[86]S[89]T[93]K[97]
575.819	38.50	SALVTVSSASTK	LVTVSSASTK	S[44]A[50]L[97]V[99]T[99]V[99]S[98]S[96]A[85]S[88]T[90]K[95]
575.819	38.55	TGLVTVSSASTK	LVTVSSASTK	T[46]G[57]L[98]V[99]T[99]V[99]S[98]S[96]A[86]S[90]T[93]K[97]
575.819	38.59	TGLVTVSSASTK	LVTVSSASTK	T[46]G[52]L[98]V[99]T[99]V[99]S[99]S[97]A[87]S[89]T[93]K[97]
575.822	38.40	SALVTVSSASTK	LVTVSSASTK	S[61]A[50]L[97]V[99]T[99]V[99]S[96]S[89]A[82]S[89]T[94]K[99]
575.822	38.45	SALVTVSSASTK	LVTVSSASTK	S[44]A[51]L[97]V[99]T[99]V[99]S[98]S[96]A[85]S[88]T[93]K[98]
575.822	38.55	SALVTVSSASTK	LVTVSSASTK	S[43]A[51]L[98]V[99]T[99]V[99]S[97]S[92]A[85]S[89]T[94]K[98]
575.822	38.59	TGLVTVSSASTK	LVTVSSASTK	T[69]G[71]L[98]V[99]T[99]V[99]S[98]S[95]A[83]S[90]T[95]K[97]
575.822	38.50	TGLVTVSSASTK	LVTVSSASTK	T[55]G[62]L[97]V[99]T[99]V[99]S[98]S[96]A[86]S[89]T[92]K[95]

The PEAKS software provides an integrated filtering option through ALC value and is applicable to each of the PEAKS analysis files. Although selection by ALC is simple to implement, it is not robust in extracting useful data. In particular, short peptide assignments with very poor local confidences at a few residues will confer a low ALC value even if there exists a contained subsequence of high confidence. Further still, peptides that have an ALC above the filter threshold may have regions of poor local confidence that can misrepresent the data. The method employed by the IgFamily program is to retrieve high local confidence subsequences by use of a rolling average filter, requiring any 3 (less at a terminal) contiguous residues to have an average local confidence greater than a defined value.

The example in Table 1 shows peptides filtered with a rolling average method. Inspecting the peptide assignments of SALVTVSSASTK and TGLVTVSSASTK reveals similar precursor mass-to-charge ratios and retention times, suggesting that these sequences were likely produced by the same originating peptide. The differing factor here is poor N-terminal local confidences. Applying a rolling average method effectively truncates the areas of low local confidence and produces sequences of high quality.

- Data: what does it mean?

There is additional knowledge in the data to give evidence to the likelihood of assigning a peptide to a protein. A protein that has accrued substantial evidence should influence the probability of assigning further peptides to that protein. This *prior evidence* is an important consideration when viewing the population of peptides in the sample - If knowledge exists of the likely peptide distribution on the basis of protein abundance *a priori*, this should be represented in the conditional probability. How is this holistic view achievable? The answer lies in the creation of a robust statistical model.

- There's method in models

Sequence similarity is a routine measure for the likelihood of a peptide assignment to a protein and is most often referred to as the *homology* of a peptide $\boldsymbol{p}$ for a protein $\boldsymbol{g}$. The determination of a homology value requires a defined rate of amino acid substitution. Although the substitution itself occurs at the nucleotide level, some observed products are more favourable as a result of the physical likelihood of a residue substitution and the likelihood of the protein retaining its function (at least if it is necessary for the fitness of the organism). This can be determined by considering the phylogenetic frequency of substitution in a conserved region of a protein and creating a *substitution matrix*. The substitution matrix used for the IgFamily program is the BLOSUM62 matrix generated by comparison of transmembrane proteins (Henikoff & Henikoff, 1992) (Figure 2). Although routinely used for sequence similarity, it may not necessarily represent the rate of substitution in immunoglobulins. This is examined further in the Considerations chapter.

If homology is considered as a distance, with peptide sequences to a more similar protein sequence being spatially 'closer', then the homology of a peptide sequence is approximated by a *Poisson distribution* (Karlin & Altschul, 1990) (Figure 3). The homology of a peptide $\boldsymbol{p}$ for a protein $\boldsymbol{g}$ can be viewed as a relative value against all other homologies possible for any collection of proteins $\boldsymbol{G}$. This value is defined as the homology density $\boldsymbol{D_g(p)}$ of a peptide $\boldsymbol{p}$ for a protein $\boldsymbol{g}$. The value of $\boldsymbol{D_g(p)}$ must be inclusively between 0 and 1. A homology density $\boldsymbol{D_g(p)}$ close to 0 states that the homology of the peptide $\boldsymbol{p}$ for a protein $\boldsymbol{g}$ is a minor value compared to all the homologies possible for the peptide $\boldsymbol{p}$ for any protein in $\boldsymbol{G}$. A high value of the homology density $\boldsymbol{D_g(p)}$ states that the homology of the peptide $\boldsymbol{p}$ for a protein $\boldsymbol{g}$ is a large proportion of all possible homology values. In particular, a peptide with a large density value for a protein is defined as a *distinct peptide* for that protein. Distinct peptides are most likely derived from a single or a few proteins and as such the majority of the assignment likelihood is centred on those proteins, represented in Figure 3 by a low Poisson rate factor. In contrast, peptides that are not resolvable from many proteins have a shared assignment likelihood and a high Poisson rate factor.

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Figure 2 - The BLOSUM62 substitution matrix (Henikoff & Henikoff, 1992). Although routinely used for sequence similarity, it may not necessarily represent the rate of substitution for immunoglobulins.

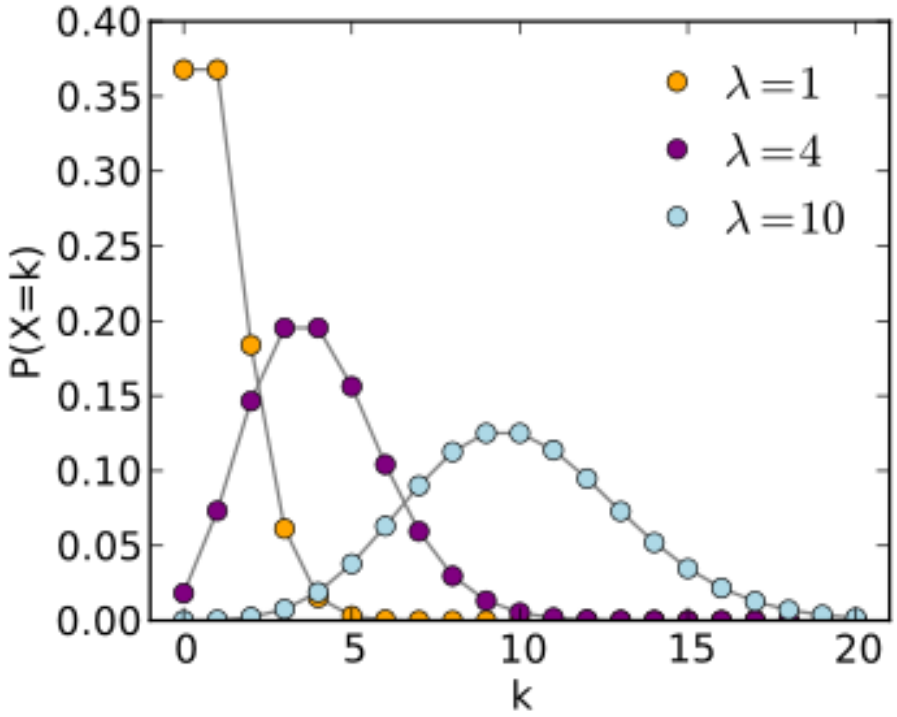


Figure 3 - A Poisson distribution showing three scaling factors. Peptides with a high homology density are likely to have a small scaling factor.

The BLAST (Basic Local Alignment Search Tool) program is a utility for realising Poisson modelled assignment likelihood of a peptide query in comparison to a protein subject database (Altschul, et al., 1990). Although useful as an aid for peptide or protein identification, the BLAST algorithm is limited in its ability to consider the evidence of a sample that may be known or determinable. As an example, consider a sample where many peptides have already been analysed. Suppose that the majority of those peptides are distinct for the gene family IGHV1-69. It is intuitive that further peptides for this sample should be more likely to be assigned to IGHV1-69. Distinct peptides may receive only a smaller relative increase in assignment likelihood, by nature of the assignment being near-certain initially. Those peptides that are potentially shared among a few peptides will benefit from the evidence contained in the entirety of the sample. Peptides that are shared may need a greater amount of evidence to escape the constraints of a relatively weak assignment. In particular, peptides that are heavily shared may never be confidently assigned as the assignment uncertainty could be greater than the resolving power of the data.

The standard method for adjustment of the peptide assignment Poisson model is by conjunction with a secondary model on the basis of prior or determined evidence. Although in general any model has functionality to transform a Poisson distribution, some models are more suitable by their ability to simplify the conjugation. These models are known as *conjugate models* or *conjugate priors*. An appropriate conjugate prior for the Poisson distribution is an exponential distribution (Figure 4).

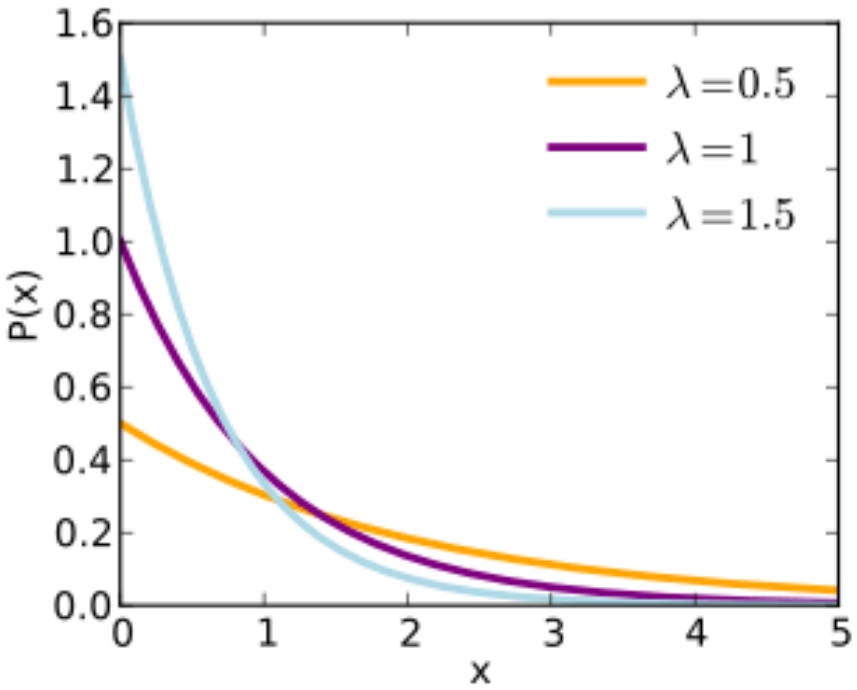


Figure 4 - An exponential distribution showing three scaling factors. The scaling factor represents how strongly prior or determined evidence is considered.

The Poisson-exponential conjugation is sufficient to represent peptide assignment with consideration of evidence determinable from the sample, but how is the evidence itself represented? The canonical model for describing elements that originate from one of many categories is the multinomial distribution. It is clear that any peptide that is observed can only physically result from a single protein. However, the ontological question is how likely such a peptide is a result from one of potentially many proteins. The multinomial distribution allows insight from several questions that take the form of the conditional probability $\boldsymbol{p}(\boldsymbol{g} \mid \boldsymbol{p})$ -

- For any peptide $\boldsymbol{p}$, without knowing the peptide sequence, what is the likelihood that it belongs to some protein $\boldsymbol{g}$?

- For any peptide $\boldsymbol{p}$, knowing the peptide sequence, what is the likelihood that it belongs to some protein $\boldsymbol{g}$?

- Without knowing the origin protein $\boldsymbol{g}$, what is the likelihood of observing a peptide $\boldsymbol{p}$?

- Knowing the origin protein $\boldsymbol{g}$, what is the likelihood of observing a peptide $\boldsymbol{p}$?

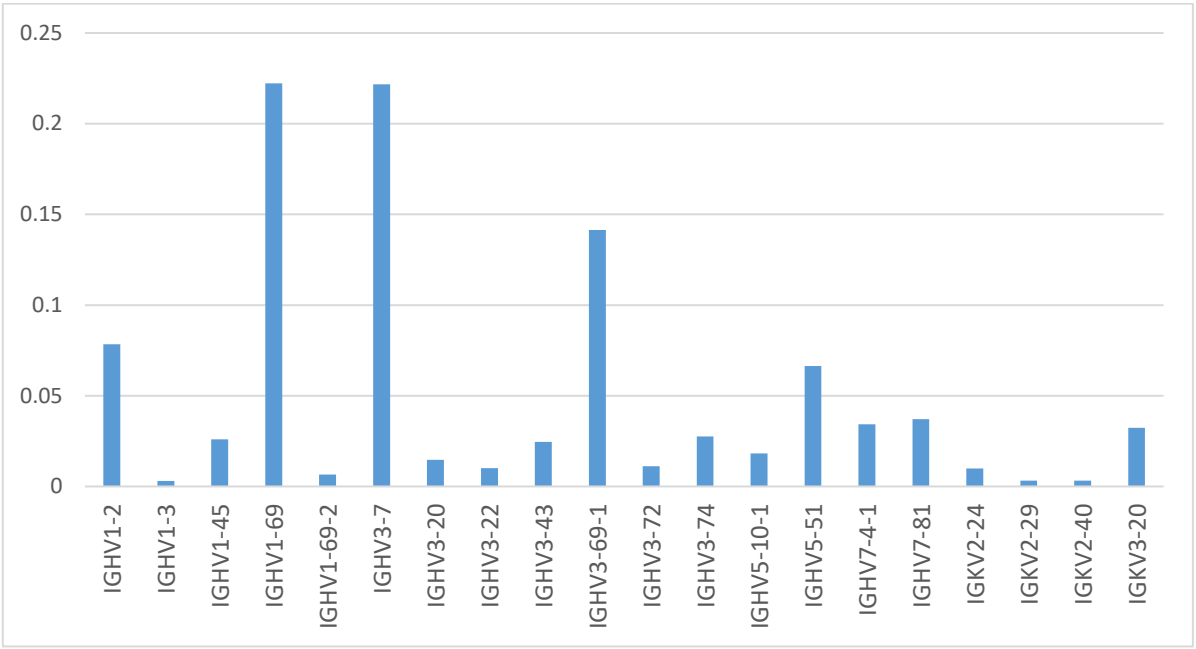


Figure 5 - A multinomial distribution for categorical variables. Here the categories are the twenty most likely gene families for the sample WM16_B02+B02a_HAGG_isolated_RF_MJ2. The x-axis shows the gene family and the y-axis shows probability density of the conditional probability $\boldsymbol{p}(\boldsymbol{g} \mid \boldsymbol{p})$.

The value of multinomial categories is typically the product of a likelihood function that assigns a likelihood value for each peptide through an appropriate model. The model used by the IgFamily is the following:

First, determine the homology $\boldsymbol{H_g(p)}$ of the peptide $\boldsymbol{p}$ for the protein $\boldsymbol{g}$, as sampled from a Poisson distribution, and repeat for all peptides $\boldsymbol{p} \in \boldsymbol{P}$ and proteins $\boldsymbol{g} \in \boldsymbol{G}$. Note that each of all peptides have a strictly positive (although possibly insignificant) homology for each of all proteins -

$$\boldsymbol{H_g(p) \sim Poisson} \; \forall \{\boldsymbol{p \in P, g \in G}\}$$

Determine the transformed homology $\boldsymbol{T_g(p)}$ of the peptide $\boldsymbol{p}$ for the protein $\boldsymbol{g}$ and repeat for all peptides $\boldsymbol{p} \in \boldsymbol{P}$ and proteins $\boldsymbol{g} \in \boldsymbol{G}$. The transformed homology modifies the homology $\boldsymbol{H_g(p)}$ through a constant scalar factor $\boldsymbol{\alpha_1}$ that represents the weight of relative homology values (a difference between a large and a small value becomes greater) and a factor $\boldsymbol{\alpha_2}$ that penalises the homology by the quantity of mismatched amino acids $\boldsymbol{E_g(p)}$.

$$\boldsymbol{T_g(p) = H_g(p)^{\alpha_1} \cdot \alpha_2^{E_g(p)}}$$

Determine the conjugated transformed homology $\boldsymbol{Z_g(p)}$ of the peptide $\boldsymbol{p}$ for the protein $\boldsymbol{g}$ and repeat for all peptides $\boldsymbol{p} \in \boldsymbol{P}$ and proteins $\boldsymbol{g} \in \boldsymbol{G}$. The conjugated transformed homology is the product of the transformed homology $\boldsymbol{T_g(p)}$ with the conjugation value $\left(\frac{\boldsymbol{M(g))}}{\boldsymbol{max(M)}}\right)^{\boldsymbol{\beta}}$. The conjugation value is the value of the protein $\boldsymbol{g}$ in proportion to the maximum scoring protein, raised to the power of a non-negative scalar $\boldsymbol{\beta}$. Notice that $\boldsymbol{Z_g(p)}$ is dependent on $\boldsymbol{S_g(p)}$ which is itself dependent on $\boldsymbol{Z_g(p)}$ through $\boldsymbol{C_g(p)}$. As described shortly, initially $\boldsymbol{\beta = 0}$, such that $\boldsymbol{Z_g(p) = T_g(p)}$, with $\boldsymbol{\beta}$ increasing as the algorithm converges towards a clustering of gene families -

$$\boldsymbol{Z_g(p) = T_g(p) \cdot \left(\frac{M(g)}{max(M)}\right)^{\beta} \; \forall \{p \in P, g \in G\} \; \forall ! \{\beta \geq 0\}}$$

Determine the transformed homology density $\boldsymbol{D_g(p)}$ of the peptide $\boldsymbol{p}$ for the protein $\boldsymbol{g}$ and repeat for all peptides $\boldsymbol{p} \in \boldsymbol{P}$ and proteins $\boldsymbol{g} \in \boldsymbol{G}$. The transformed homology density is the value of the transformed homology $\boldsymbol{T_g(p)}$ of the peptide $\boldsymbol{p}$ for the protein $\boldsymbol{g}$ in comparison to the sum $\boldsymbol{\sum_{g \in G} T_g(p)}$ of all transformed homologies for the peptide $\boldsymbol{p}$ for all $\boldsymbol{g} \in \boldsymbol{G}$ proteins -

$$\boldsymbol{D_g(p) = \frac{T_g(p)}{\sum_{g \in G} T_g(p)} \; \forall \{p \in P, g \in G\}}$$

Determine the conjugated homology density $\boldsymbol{C_g(p)}$ of the peptide $\boldsymbol{p}$ for the protein $\boldsymbol{g}$ and repeat for all peptides $\boldsymbol{p} \in \boldsymbol{P}$ and proteins $\boldsymbol{g} \in \boldsymbol{G}$. The conjugated homology density is the value of the conjugated homology $\boldsymbol{Z_g(p)}$ of the peptide $\boldsymbol{p}$ for the protein $\boldsymbol{g}$ in comparison to the sum $\boldsymbol{\sum_{g \in G} Z_g(p)}$ of all conjugated homologies for the peptide $\boldsymbol{p}$ for all $\boldsymbol{g} \in \boldsymbol{G}$ proteins -

$$\boldsymbol{C_g(p) = \frac{Z_g(p)}{\sum_{g \in G} Z_g(p)} \; \forall \{p \in P, g \in G\}}$$

Determine the peptide score $\boldsymbol{S_g(p)}$ of the peptide $\boldsymbol{p}$ for the protein $\boldsymbol{g}$ and repeat for all peptides $\boldsymbol{p} \in \boldsymbol{P}$ and proteins $\boldsymbol{g} \in \boldsymbol{G}$. The peptide score is a product of the conjugated homology density $\boldsymbol{C_g(p)}$ and the transformed homology density $\boldsymbol{D_g(p)}$ raised to the power of a non-negative scalar $\boldsymbol{\gamma}$. The product represents the Poisson-exponential conjugation, with $\boldsymbol{D_g(p)}$ functionalising the Poisson and $\boldsymbol{C_g(p)}$ functionalising the exponential. The scalar $\boldsymbol{\gamma}$ provides weighting to the prior distribution of the Poisson and represents the naivety of the model. A greater $\boldsymbol{\gamma}$ will result in values of $\boldsymbol{S_g(p)}$ that do not diverge greatly from the initially assigned transformed homology density $\boldsymbol{D_g(p)}$. In constrast, as evidence towards the conjugated homology density $\boldsymbol{C_g(p)}$ increases, most notably with a greater sample size (a larger count of observed peptides), $\boldsymbol{\gamma}$ will approach zero and the peptide score $\boldsymbol{S_g(p)}$ will be dominated by $\boldsymbol{C_g(p)}$ -

$$\boldsymbol{S_g(p) = C_g(p) \cdot D_g(p)^{\gamma} \; \forall \{p \in P, g \in G\}}$$

Determine the multinomial value $\boldsymbol{M(g)}$ for the protein $\boldsymbol{g}$ and repeat for all proteins $\boldsymbol{g \in G}$. The multinomial value is the summation of all the peptide scores $\boldsymbol{\sum_{p \in P} S_g(p)}$ for all peptides $\boldsymbol{p \in P}$. The multinomial value represents the evidence for the protein as inferred by the model.

$$\boldsymbol{M(g) = \sum_{p \in P} S_g(p)}$$

Finally, determine the multinomial density $\boldsymbol{p(g)}$ for the protein $\boldsymbol{g}$ and repeat for all proteins $\boldsymbol{g \in G}$. The multinomial density is the value of the multinomial value $\boldsymbol{M(g)}$ for the protein $\boldsymbol{g}$ in comparison to the sum $\boldsymbol{\sum_{g \in G} M(g)}$ of all multinomial values -

$$\boldsymbol{p(g) = \frac{M(g)}{\sum_{g \in G} M(g)} \; \forall \{g \in G\}}$$

The multinomial density is an important measure of inference. In particular, it reflects the question "If any single gene family protein were observed from a sample of gene family proteins, what is the likelihood that it is gene family $\boldsymbol{g}$?". This value also determines the proportion of gene families present in the sample. Note that the sum of multinomial densities is equal to 1 - This is typical of a full probability distribution (the combination of all possible outcomes in the model is representative of every possible outcome).

There is some depth in defining a clustering value for $\boldsymbol{\beta}$, however it is not discussed at this time. The general principle behind clustering in this way presumes that there is a reasonable belief in the definite amount of gene families that are present in a sample. The IgFamily program clusters on a conservative basis that at least 80% of the multinomial value should be contained in the top 7 gene families. An optimal solution to clustering can be determined through the principle of *maximum entropy* and is retained for future work.

Function

The IgFamily program is a console based application developed using the C++ programming language. It operates by reading export files produced by the PEAKS database, PEAKS de novo, or NOVOR de novo programs along with a FASTA file containing a protein database. These upstream programs create a peptide assignment for each of the spectra contained in a mass spectrometry data file. Through this, the IgFamily program creates a data structure of peptide assignments and a reference data structure of proteins and determines the association of these peptides to the proteins by the model described in the Story chapter. The initial homology values for peptide assignment to proteins is determined by the BLAST program which is externally integrated into the IgFamily program.

Functionality is modifiable by the user through a console-based interface. The user is able to select which FASTA file(s) to use, create a custom FASTA file, select a peptide assignment method, and include the msconvert file conversion and NOVOR de novo programs into the workflow. The resulting output is designed around two HTML files that display the peptide assignments and most probable gene families. A description of key features follows and concludes with the pseudocode of the IgFamily program.

- Cross platform compatibility

The default version of the IgFamily requires a 64-bit Windows operating system and may be compiled for 32-bit Windows, 32-bit and 64-bit Mac OS X, and 32-bit and 64-bit Linux distributions.

- FASTA utility tools

A subroutine is able to parse FASTA files into runtime. The input format is standardised as:

>[ACCESSION]|[NAME]|[TYPE]|[SPECIES]|

A FASTA creation utility provides functionality for creating custom FASTA files with any combination of fields. This utility can be used to create the required FASTA format and may be used to include parsing rules for any input format. Accession field is an identifier field and is not required for runtime. Name, type, and species fields are used in data structure creation, association, and function.

- Peptide file compatibility

Existing software is able to determine a peptide assignment for generated spectra. The IgFamily program is able to parse export files created by these programs for downstream analysis. The currently supported software are the PEAKS v8.0 DE NOVO de novo peptides .csv export, the PEAKS v8.0 SPIDER protein-peptide .csv export, and the NOVOR v1.1 de novo peptides .csv export.

For the PEAKS v8.0 DE NOVO de novo peptides export file IgFamily parses the scan number, peptide accession (identifier value), peptide assignment with modification, peptide mass-to-charge ratio, peptide retention time, peptide theoretical mass, peptide amino acid local confidence score, and source file name. Note that PEAKS v8.0 DE NOVO assigns individual peptide accessions for replicate peptide assignments.

For the PEAKS v8.0 SPIDER protein-peptide export file IgFamily parses the scan number, peptide accession (identifier value), peptide assignment with modification, peptide mass-to-charge ratio, peptide retention time, peptide theoretical mass, peptide spectral count, peptide assignment confidence value IgP, and source file name.

For the NOVOR v1.1 de novo peptides file export file IgFamily parses the scan number, peptide accession (identifier value), peptide with modification, peptide mass-to-charge ratio, peptide retention time, peptide theoretical mass, peptide amino acid local confidence score, and source file name. Note that NOVOR v1.1 assigns individual peptide accessions for replicate peptide assignments.

- msconvert external integration

The msconvert mass spectrometry data file convertor is able to be called through a user defined interface option to convert AB Sciex .wiff and .wiff.scan file types into the Mascot Generic Format .mgf file type. Various command line options may be selected with peak-picking as the default option. The generated file is created in the same folder as the input file.

- NOVOR v1.1 external integration

The NOVOR v1.1 de novo command line application is able to be called through a user defined interface option to generate NOVOR v1.1 de novo peptide assignment files of the .csv file type. Various command line options may be selected. The generated file is created in the same folder as the input file.

- Local directory functionality

The IgFamily program can be run in local directory or filesystem directory file mode. In local directory file mode the user places the required files in the IgFamily root directory and executes the program. The input file and output files are moved to a folder created in a subdirectory with the name of the input file. The local file mode is able to be executed from any directory.

- Filesystem functionality

The IgFamily program is able to be run in local directory or filesystem mode. In filesystem mode the program accessions a dedicated file association structure to retrieve and export files. The filesystem is currently defined on the *FATELVIS* network assisted storage device. Initially the user is required to accession an input file (Figure 6), however there is proposed functionality for dynamic file management with the ISO/IEC TS 18822:2015 filesystem library. A useful feature of the filesystem is the addition of data in an accession file not also included in the input files. In the example in Figure 6 the patient status is primary Sjögren's Syndrome (pSS) - With a large scale study this would allow factor analysis between files with different covariates: patient status, progression of disease at sampling, treatment, mass spectrometry settings, and any other variables of interest.

```
ID: WM16_B02+B02a_HAGG_isolated_RF_MJ2;
FILE: WM16_B02+B02a_HAGG_isolated_RF_MJ2;
VERSION: v0.12.3;
STATUS: pSS;
ENZYME: Trypsin;
DENOVO_DELTAMASS: 0.02;
```

Figure 6 - An example of a filesystem accession file. Currently the user is required to create an accession file to operate the IgFamily program in filesystem mode.

- Runtime user interface

At initialisation of the IgFamily program the user is greeted with an interactive menu. The user is able to access FASTA file utilities, msconvert command line tools, NOVOR v1.1 command line tools, and select program parameters. Program parameters include local or filesystem file modes, FASTA file selection, and peptide assignment selection.

```
Current settings:

FASTA file -                    IGHV_IGLV_IGKV_CONT_UNIPROT_20160827.fasta
Spectra assignment method -     PEAKS de novo

Select FASTA file:                          [F]
Select spectra assignment method:           [P]
Continue with current settings:             [X]

--> F


Current setting - IGHV_IGLV_IGKV_CONT_UNIPROT_20160827.fasta

[0] IGHV_IGLV_IGKV_20160827.fasta
[1] IGHV_IGLV_IGKV_CONT_20160827.fasta
[2] IGHV_IGLV_IGKV_CONT_UNIPROT_20160827.fasta
[3] IGHV_IGLV_IGKV_mABV_CONT_20160827.fasta
[4] IGHV_IGLV_IGKV_MIGHV_MIGKV_MIGLV_20160827.fasta
[5] IGHV_IGKV_IGLV_MIGHV_MIGKV_MIGLV_CONT_20160827.fasta
[6] IGH_IGL_IGK_20160827.fasta
[7] IGH_IGL_IGK_CONT_20160827.fasta
[8] IGH_IGL_IGK_CONT_UNIPROT_20160827.fasta
[9] IGH_IGL_IGK_mABV_CONT_20160827.fasta
[X] Use current setting

-->
```

Figure 7 - The user can select a FASTA file through the interface.

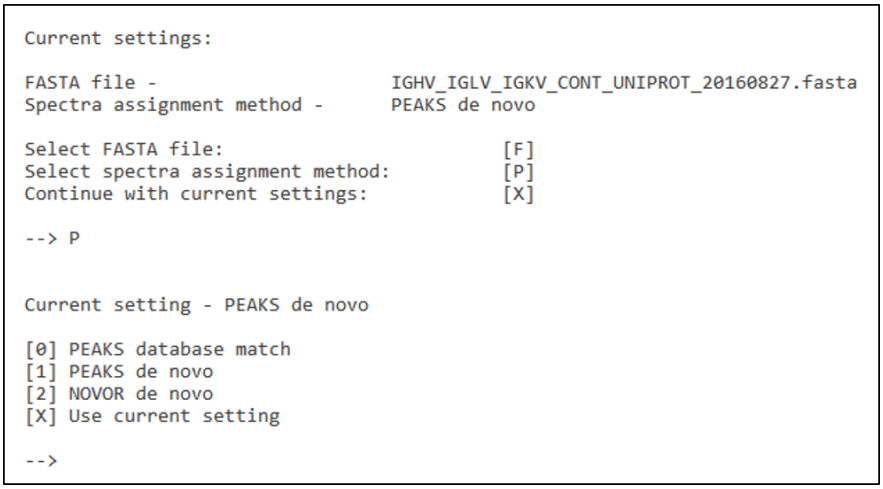


Figure 8 - The user can select a peptide assignment method through the interface.

- FASTA data structuring

Following user selection of runtime parameters, the IgFamily program will parse the FASTA file(s). The [NAME], [TYPE], and [SPECIES] fields have runtime functionality, although the [ACCESSION] field is retained for FASTA utility functions. An example excerpt is shown in Table 2.

- Peptide data structuring

Following user selection of runtime parameters, the IgFamily program will parse the data .csv file(s). From the data, the raw peptide creates two additional data types for the peptide with modifications removed and the peptide truncated based on associated de novo local confidence scores (Table 1). Truncation requires a moving average of 85% for amino acid local confidence and a minimum peptide length of 5. In the event that a peptide is cleaved at a midpoint such that two peptides of length >= 5 are produced both are assigned to unique data structures. An example excerpt is shown in Table 3.

- Homology data structuring

Following data structuring of FASTA and peptide data, a BLAST reference database is created from the FASTA data and an input peptide query list is created from the peptide data. The input peptide queries are measured for similarity against the BLAST reference database and a results file is generated. BLAST is programmed to allow up to 200 matches for each query, and a generous threshold for alignment acceptance. The BLAST output file is parsed into runtime and peptide queries are associated to their peptide data structure and protein data structure counterparts. With the peptides associated they can be scored through the model proposed in the Story chapter. An example excerpt is shown in Table 4.

- Protein scoring

With the homology data structure created, peptides determine the protein scoring as proposed in the Story chapter. An example excerpt is shown in Table 5.

- Probable germline determination

With the homology data assigned and proteins scored the IgFamily program determines the most likely germline allele usage and uses this to a refined BLAST database. The IgFamily program assumes a single allele of each gene family is used. This is followed with a second iteration of homology scoring and data structuring.

- Model conjugation

To consider the overall evidence contained in the sample the IgFamily program transforms the initial peptide homology values with a second conjugate function as proposed in Story chapter. Adjusted homology values modify peptide scores and consequent protein scores. The conjugation process is iterated until the clustering condition is achieved.

- Report generation

As a result of its analysis the IgFamily program produces two primary HTML reports: a summary report showing the most likely present gene families with consensus results and an expanded report that additionally includes the assigned peptides. The gene families are ranked by the multinomial value $\boldsymbol{M}(\boldsymbol{g})$ while the peptides are ranked by the peptide score $\boldsymbol{S_g}(\boldsymbol{p})$**.** The consensus data displayed at the header of each protein are the highest scoring peptides asssigned to the protein. In the following pages the summary report is shown along with the expanded report for the top three gene families. Both reports display the consensus information. Note that all proteins in the database are considered while only the gene family proteins are shown. The HTML report includes colour-coding for fast readability (Table 7). The colour-coding for peptides in the expanded peptide report are the local confidence values. For database matched peptides these are displayed as if they were assigned with 100% de novo certainty.

Table 2 - Example excerpt of protein_data report file.

key	protein_name	protein_type	protein_species	protein_protein
0	IGHV1-18*01	IGV	Homo_sapiens	QVQLVQSGAEVKKPGASVKVSCKASGYTFTSYGISWVRQAPGQGLEWMGWISAYNGNTNYAQKLQGRVTMTTDTSTSTAYMELRSLRSDDTAVYYCAR
1	IGHV1-18*02	IGV	Homo_sapiens	QVQLVQSGAEVKKPGASVKVSCKASGYTFTSYGISWVRQAPGQGLEWMGWISAYNGNTNYAQKLQGRVTMTTDTSTSTAYMELRSLRSDDTA
2	IGHV1-18*03	IGV	Homo_sapiens	QVQLVQSGAEVKKPGASVKVSCKASGYTFTSYGISWVRQAPGQGLEWMGWISAYNGNTNYAQKLQGRVTMTTDTSTSTAYMELRSLRSDDMAVYYCAR
3	IGHV1-18*04	IGV	Homo_sapiens	QVQLVQSGAEVKKPGASVKVSCKASGYTFTSYGISWVRQAPGQGLEWMGWISAYNGNTNYAQKLQGRVTMTTDTSTSTAYMELRSLRSDDTAVYYCAR
4	IGHV1-2*01	IGV	Homo_sapiens	QVQLVQSGAEVKKPGASVKVSCKASGYTFTGYYMHWVRQAPGQGLEWMGRINPNSGGTNYAQKFQGRVTSTRDTSISTAYMELSRLRSDDTVVYYCAR
5	IGHV1-2*02	IGV	Homo_sapiens	QVQLVQSGAEVKKPGASVKVSCKASGYTFTGYYMHWVRQAPGQGLEWMGWINPNSGGTNYAQKFQGRVTMTRDTSISTAYMELSRLRSDDTAVYYCAR
6	IGHV1-2*03	IGV	Homo_sapiens	QVQLVQSGAEVKKLGASVKVSCKASGYTFTGYYMHWVXQAPGQGLEWMGWINPNSGGTNYAQKFQGRVTMTRDTSISTAYMELSRLRSDDTAVYYCAR
7	IGHV1-2*04	IGV	Homo_sapiens	QVQLVQSGAEVKKPGASVKVSCKASGYTFTGYYMHWVRQAPGQGLEWMGWINPNSGGTNYAQKFQGWVTMTRDTSISTAYMELSRLRSDDTAVYYCAR
8	IGHV1-2*05	IGV	Homo_sapiens	QVQLVQSGAEVKKPGASVKVSCKASGYTFTGYYMHWVRQAPGQGLEWMGRINPNSGGTNYAQKFQGRVTMTRDTSISTAYMELSRLRSDDTVVYYCAR
9	IGHV1-24*01	IGV	Homo_sapiens	QVQLVQSGAEVKKPGASVKVSCKVSGYTLTELSMHWVRQAPGKGLEWMGGFDPEDGETIYAQKFQGRVTMTEDTSTDTAYMELSSLRSEDTAVYYCAT

Table 3 - Example excerpt of peptide_data output file.

key	scan_ID	peptide_mz	peptide_z	peptide_rt	peptide_m	peptide_withmod	peptide_withoutmod	peptide_filtered
1036	19205	800.905	2	41.85	1599.81	A(+27.99)APSGVTTDKVQAEAK	AAPSGVTTDKVQAEAK	SGVTTDKVQAEAK
1675	19035	800.906	2	41.33	1599.81	A(+27.99)APSGVTTDKVQAEAK	AAPSGVTTDKVQAEAK	VTTDK
3559	27064	696.285	3	38.96	2085.85	A(+27.99)CSVSCGQ(+.98)LCDLLECKDDR	ACSVSCGQLCDLLECKDDR	QLCDLL
1025	17823	1068.51	2	36.06	2135.00	A(+27.99)DNALKSGNSKESVTEQDSK	ADNALKSGNSKESVTEQDSK	ALKSGNSKESVTEQ
3420	26503	719.986	3	36.31	2156.95	A(+27.99)DNALQSYMNEVSTEQTTK	ADNALQSYMNEVSTEQTTK	ADNALQ
1296	16439	447.714	2	30.05	893.413	A(+27.99)EGPTTYK	AEGPTTYK	GPTTYK
1523	18849	659.798	2	40.63	1317.58	A(+27.99)ESTAVCLEDPK(+27.99)	AESTAVCLEDPK	AESTAV
3234	27467	659.790	2	40.73	1317.58	A(+27.99)ESTAVLCEDPK(+27.99)	AESTAVLCEDPK	ESTAV
911	17579	809.418	2	34.83	1616.84	A(+27.99)KGSGVTTDKVQAEAK	AKGSGVTTDKVQAEAK	GSGVTTDKVQAEAK
1104	17888	809.417	2	36.28	1616.84	A(+27.99)KGSGVTTDKVQAEAK	AKGSGVTTDKVQAEAK	GSGVTTDKVQAEAK

denovo_peptide	denovo_peptide_filtered
A[39]A[49]P[37]S[94]G[95]V[98]T[99]T[98]D[98]K[95]V[94]Q[92]A[96]E[98]A[93]K[97]	S[94]G[95]V[98]T[99]T[98]D[98]K[95]V[94]Q[92]A[96]E[98]A[93]K[97]
A[27]A[31]P[18]S[44]G[75]V[88]T[96]T[89]D[90]K[86]V[81]Q[76]A[82]E[92]A[80]K[91]	V[88]T[96]T[89]D[90]K[86]
A[71]C[28]S[26]V[60]S[56]C[20]G[11]Q[87]L[93]C[89]D[93]L[94]L[93]E[86]C[11]K[10]D[56]D[79]R[93]	Q[87]L[93]C[89]D[93]L[94]L[93]
A[46]D[61]N[76]A[87]L[95]K[97]S[98]G[91]N[93]S[95]K[93]E[97]S[89]V[91]T[92]E[95]Q[79]D[93]S[65]K[86]	A[87]L[95]K[97]S[98]G[91]N[93]S[95]K[93]E[97]S[89]V[91]T[92]E[95]Q[79]
A[86]D[95]N[95]A[96]L[98]Q[95]S[87]Y[31]M[11]N[87]E[96]V[81]S[72]T[82]E[87]Q[65]T[18]T[24]K[62]	A[86]D[95]N[95]A[96]L[98]Q[95]
A[32]E[57]G[92]P[90]T[94]T[94]Y[92]K[99]	G[92]P[90]T[94]T[94]Y[92]K[99]
A[86]E[93]S[85]T[95]A[95]V[93]C[76]L[79]E[86]D[69]P[24]K[29]	A[86]E[93]S[85]T[95]A[95]V[93]
A[82]E[92]S[85]T[96]A[97]V[93]L[66]C[43]E[82]D[85]P[51]K[57]	E[92]S[85]T[96]A[97]V[93]
A[35]K[43]G[94]S[98]G[94]V[97]T[99]T[98]D[98]K[95]V[95]Q[95]A[92]E[96]A[92]K[97]	G[94]S[98]G[94]V[97]T[99]T[98]D[98]K[95]V[95]Q[95]A[92]E[96]A[92]K[97]
A[37]K[41]G[90]S[97]G[85]V[93]T[98]T[91]D[93]K[89]V[95]Q[84]A[83]E[91]A[87]K[92]	G[90]S[97]G[85]V[93]T[98]T[91]D[93]K[89]V[95]Q[84]A[83]E[91]A[87]K[92]

Table 4 - Example excerpt of homology_data report file.

key_query	query	subject	key_subject_accession	subject_accession	replicate_count	mismatch_count	alignment_coverage_delta	alignment_coverage
17	YYVDSVK	YYVDSVK	184	IGHV3-7	27	0	0	100
17	YYVDSVK	YYVDSVK	160	IGHV3-52	27	0	0	100
17	YYVDSVK	HYVDSVK	82	IGHV3-16	27	1	0	100
17	YYVDSVK	YYADSVK	165	IGHV3-53	27	1	0	100
17	YYVDSVK	YYADSVK	178	IGHV3-66	27	1	0	100
17	YYVDSVK	YYADSVK	96	IGHV3-23	27	1	0	100
17	YYVDSVK	YYADSVK	147	IGHV3-43	27	1	0	100
17	YYVDSVK	YYADSVK	131	IGHV3-30-5	27	1	0	100
17	YYVDSVK	YYADSVK	105	IGHV3-30	27	1	0	100
17	YYVDSVK	YYADSVK	126	IGHV3-30-3	27	1	0	100

homology	homology_transformed	homology_ conjugated	homology_density_transformed	homology_density_conjugated	score
3400	2.29E+12	6.76E+11	0.47473	0.99277	26.2091
3400	2.29E+12	8.66E-67	0.47473	1.27E-78	3.36E-77
3200	2.74E+10	7.79E-130	0.00567	1.14E-141	2.88E-140
3000	2.19E+10	4.95E-07	0.00453	7.27E-19	1.74E-17
3000	2.19E+10	6.07E-08	0.00453	8.92E-20	2.14E-18
3000	2.19E+10	0.000298	0.00453	4.37E-16	1.05E-14
3000	2.19E+10	3.84E+08	0.00453	0.00057	0.01355
3000	2.19E+10	1.09E+07	0.00453	1.60E-05	0.00038
3000	2.19E+10	1.09E+07	0.00453	1.60E-05	0.00038
3000	2.19E+10	165496	0.00453	2.43E-07	5.83E-06

Table 5 - Example excerpt of protein_analysis report file.

key	protein_name	protein_score	proteinconstruct_sequencecoverage	protein_type	homology_query	homology_subject	replicate_count	peptide_score_density
85	IGHV1-69	329.172	51.02	IGV	LVQSGAEVK	LVQSGAEVK	23	0.22022
85	IGHV1-69	329.172	51.02	IGV	FGTANYAQK	FGTANYAQK	6	0.10390
85	IGHV1-69	329.172	51.02	IGV	YAISWVR	YAISWVR	7	0.09569
85	IGHV1-69	329.172	51.02	IGV	YAQNFQGR	YAQKFQGR	8	0.08158
85	IGHV1-69	329.172	51.02	IGV	VVQSGAEVK	LVQSGAEVK	7	0.05755
85	IGHV1-69	329.172	51.02	IGV	FATANYAQK	FGTANYAQK	4	0.05347
85	IGHV1-69	329.172	51.02	IGV	EDTAVY	EDTAVY	10	0.04899
85	IGHV1-69	329.172	51.02	IGV	FGTANYAQR	FGTANYAQK	3	0.04466
85	IGHV1-69	329.172	51.02	IGV	QEDTAVY	EDTAVY	11	0.04280
85	IGHV1-69	329.172	51.02	IGV	QLVQSGAEVK	QLVQSGAEVK	3	0.03137

mismatch_count	alignment_coverage_delta	homology	homology_transformed	homology_transformed_conjugated	homology_density	homology_density_conjugated
0	0	4100	4.41E+12	1.43E+12	0.09472	0.61802
0	0	4700	7.12E+12	2.31E+12	0.99999	0.99999
0	0	3500	2.54E+12	8.19E+11	0.97528	0.99998
1	0	3500	3.75E+10	1.19E+10	0.16743	0.74592
1	0	3700	4.56E+10	1.44E+10	0.08751	0.57576
1	0	3400	3.39E+10	1.07E+10	0.99999	0.99999
0	0	3000	1.48E+12	4.76E+11	0.03638	0.40311
1	0	3900	5.48E+10	1.74E+10	0.99999	0.99999
1	1	2800	1.52E+09	4.72E+08	0.02836	0.33703
0	0	4600	6.60E+12	2.14E+12	0.09436	0.61473

Table 6 - Information shown in HTML report.

Term	Model symbol	Definition
Protein	$\boldsymbol{g}$	The protein $\boldsymbol{g}$. This can be any protein from the FASTA database.
Score	$\boldsymbol{M(g)}$	The multinomial value for the protein $\boldsymbol{g}$. This value represents the total supporting evidence from peptides for the protein $\boldsymbol{g}$.
Density	$\boldsymbol{p(g)}$	The multinomial density of the gene family $\boldsymbol{g}$. This value represents the proportion of the sample that is gene family $\boldsymbol{g}$. For example a $\boldsymbol{p(g)}$ of 0.200 states that the gene family $\boldsymbol{g}$ is 20% of the gene family proteins in the sample. Note that while the model considers all proteins in the database, only the gene family proteins are considered for the mulitnomial density $\boldsymbol{p(g)}$.

Term	Model symbol	Definition
Conjugated density	$\boldsymbol{C_g(p)}$	The conjugated homology density of the peptide $\boldsymbol{p}$ for the protein $\boldsymbol{g}$. It represents the weight of the conjugated homology of the peptide $\boldsymbol{p}$ in for a particular gene family $\boldsymbol{g}$ in comparison to all other proteins it could be assigned to. For example a $\boldsymbol{C_g(p)}$ of 0.700 states that the peptide $\boldsymbol{p}$ has a 70% likelihood of originating from the protein $\boldsymbol{g}$.
Density	$\boldsymbol{D_g(p)}$	The homology density of the peptide $\boldsymbol{p}$ for the protein $\boldsymbol{g}$. It represents the weight of the homology of the peptide $\boldsymbol{p}$ in for a particular gene family $\boldsymbol{g}$ in comparison to all other proteins it could be assigned to. For example a $\boldsymbol{D_g(p)}$ of 0.400 states that the peptide $\boldsymbol{p}$ has a 40% likelihood of originating from the protein $\boldsymbol{g}$.
Conjugated homology	$\boldsymbol{Z_g(p)}$	The conjugated homology of the peptide $\boldsymbol{p}$ for the protein $\boldsymbol{g}$. The conjugated homology is the value of the homology with considering both the transformation $\boldsymbol{T_g(p)}$ and the overall evidence of the protein $\boldsymbol{M(g)}$.
Homology	$\boldsymbol{H_g(p)}$	The homology of the peptide $\boldsymbol{p}$ for the protein $\boldsymbol{g}$. This initial homology value represents the association likelihood of the peptide sequence to the database sequence. This value is produced by the BLAST sequence alignment program and is described with detail in the Story chapter.
Local confidence	$\boldsymbol{L_g(p)}$	The local confidence of the amino acid as assigned by the peptide assignment method. For database matched peptides these are displayed as if they were assigned with 100% de novo certainty.

Table 7 - The HTML report includes colour-coding for fast readability.

Term	Colour	Range
Conjugated density	BLUE	$\mathbf{1.0 \geq C_g(p) \geq 0.8}$
	GREEN	$\mathbf{0.8 > C_g(p) \geq 0.5}$
	ORANGE	$\mathbf{0.2 > C_g(p) \geq 0.0}$
	RED	$\mathbf{0.2 > C_g(p) \geq 0.0}$
Density	BLUE	$\mathbf{1.0 \geq D_g(p) \geq 0.8}$
	GREEN	$\mathbf{0.8 > D_g(p) \geq 0.5}$
	ORANGE	$\mathbf{0.2 > D_g(p) \geq 0.0}$
	RED	$\mathbf{0.2 > D_g(p) \geq 0.0}$
Conjugated homology	BLUE	$\mathbf{Z_g(p) \geq 4500}$
	GREEN	$\mathbf{4500 > Z_g(p) \geq 3500}$
	ORANGE	$\mathbf{3500 > Z_g(p) \geq 2500}$
	RED	$\mathbf{2500 > Z_g(p) \geq 0}$
Homology	BLUE	$\mathbf{H_g(p) \geq 4500}$
	GREEN	$\mathbf{4500 > H_g(p) \geq 3500}$
	ORANGE	$\mathbf{3500 > H_g(p) \geq 2500}$
	RED	$\mathbf{2500 > H_g(p) \geq 0}$
Local confidence	BLUE	$\mathbf{100 \geq H_g(p) \geq 90}$
	GREEN	$\mathbf{90 > H_g(p) \geq 80}$
	ORANGE	$\mathbf{80 > H_g(p) \geq 60}$
	RED	$\mathbf{60 > H_g(p) \geq 0}$

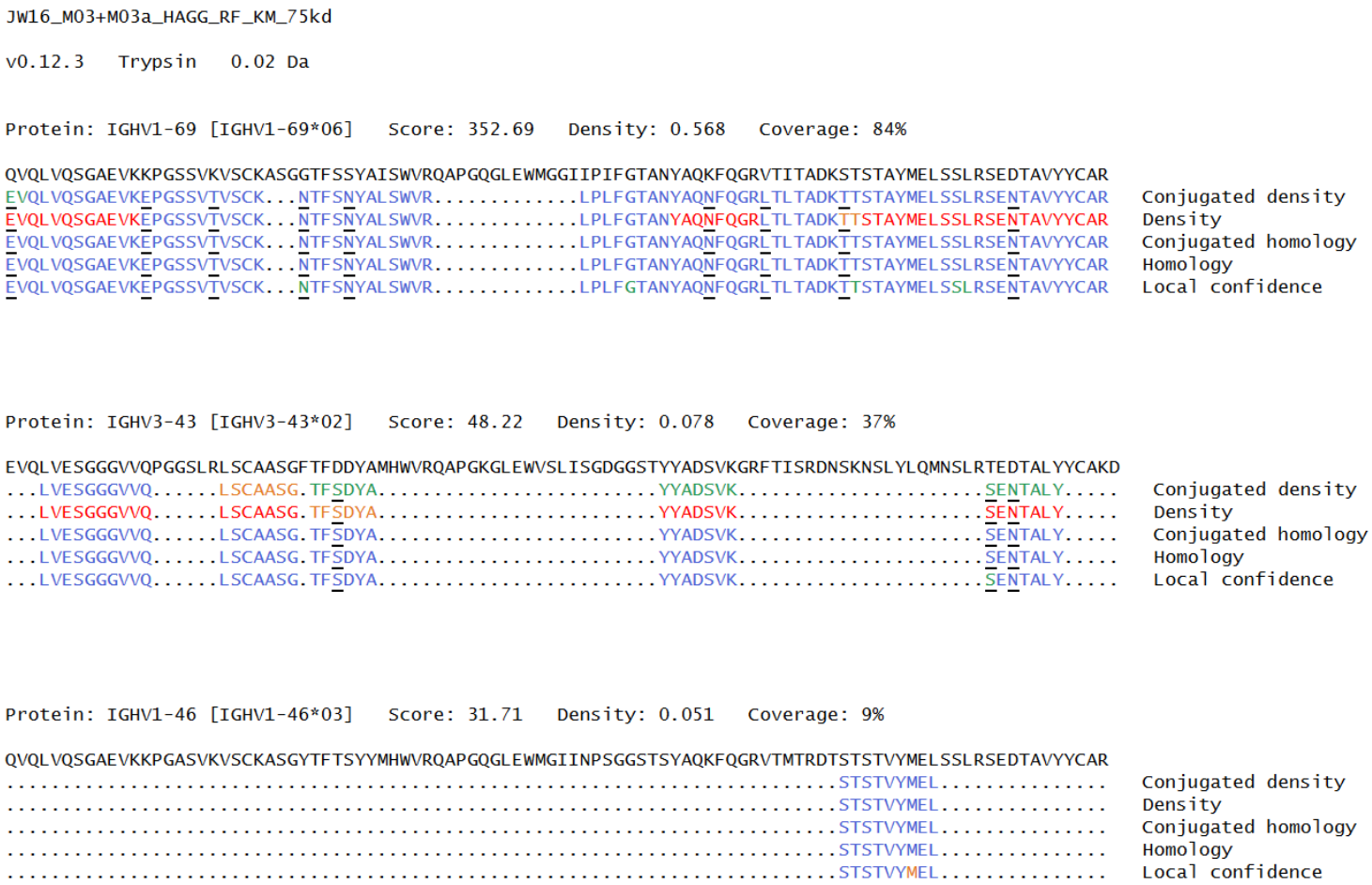


Figure 9 - Excerpt of summary HTML report for file JW16_M03+M03a_HAGG_RF_KM_75kd. The top 3 gene families are shown.

JW16_M03+M03a_HAGG_RF_KM_75kd

v0.12.3 Trypsin 0.02 Da

Protein: IGHV1-69 [IGHV1-69*06] Score: 352.69 Density: 0.568 Coverage: 84%

```
QVQLVQSGAEVKKPGSSVKVSCKASGGTFSSYAISWVRQAPGQGLEWMGGIIPIFGTANYAQKFQGRVTITADKSTSTAYMELSSLRSEDTAVYYCAR
EVQLVQSGAEVKEPGSSVTVSCK...NTFSNYALSWVR.............LPLFGTANYAQNFQGRLTLTADKTTSTAYMELSSLRSENTAVYYCAR   Conjugated density
EVQLVQSGAEVKEPGSSVTVSCK...NTFSNYALSWVR.............LPLFGTANYAQNFQGRLTLTADKTTSTAYMELSSLRSENTAVYYCAR   Density
EVQLVQSGAEVKEPGSSVTVSCK...NTFSNYALSWVR.............LPLFGTANYAQNFQGRLTLTADKTTSTAYMELSSLRSENTAVYYCAR   Conjugated homology
EVQLVQSGAEVKEPGSSVTVSCK...NTFSNYALSWVR.............LPLFGTANYAQNFQGRLTLTADKTTSTAYMELSSLRSENTAVYYCAR   Homology
EVQLVQSGAEVKEPGSSVTVSCK...NTFSNYALSWVR.............LPLFGTANYAQNFQGRLTLTADKTTSTAYMELSSLRSENTAVYYCAR   Local confidence
```

	Score	SC	Conj	Dens		
.....TFSNYAL.....	59.91	60	0.998	0.942	IGHV1-69 (1.00)	
...LVQSGAEVK.....	31.61	38	0.832	0.104	IGHV1-69 (0.83)	
.....TFSNYALSWVR.....	15.98	16	0.999	0.981	IGHV1-69 (1.00)	
...LVQSGAEVR.....	12.49	15	0.833	0.102	IGHV1-69 (0.83)	
.....YAQNFQGR.....	11.43	13	0.879	0.167	IGHV1-69 (0.88)	
.....YAQDFQGR.....	8.05	9	0.894	0.197	IGHV1-69 (0.89)	
.....LTLTADK.....	8.00	8	1.000	1.000	IGHV1-69 (1.00)	
.....SENTAVYYCAR	7.86	9	0.873	0.188	IGHV1-69 (0.87),	IGHV1-46 (0.11)
.....EPGSSVTVSCK.....	6.99	7	0.998	0.896	IGHV1-69 (1.00)	
.....YALSWVR.....	5.99	6	0.999	0.962	IGHV1-69 (1.00)	
.....STAYMELSSLR.....	5.55	6	0.924	0.131	IGHV1-69 (0.92)	
.....EDTAVY....	5.35	7	0.764	0.034	IGHV1-69 (0.76)	
.....SEDTAVY....	5.35	6	0.891	0.162	IGHV1-69 (0.89)	
.....EPGSSVTV.....	5.00	5	1.000	1.000	IGHV1-69 (1.00)	
.....FGTANYAQK.....	5.00	5	1.000	1.000	IGHV1-69 (1.00)	
.....NTFSNYAL.....	4.99	5	0.999	0.955	IGHV1-69 (1.00)	
.....SEDTALY....	4.57	6	0.762	0.090	IGHV1-69 (0.76),	IGHV3-43 (0.10)
.....STAYMEL.....	4.30	5	0.860	0.098	IGHV1-69 (0.86)	
..QLVQSGAEVK.....	4.16	5	0.833	0.103	IGHV1-69 (0.83)	
.....ANYAQNFQGR.....	4.00	4	1.000	0.999	IGHV1-69 (1.00)	
.....TFSNYALS.....	3.99	4	0.999	0.979	IGHV1-69 (1.00)	
.....STAYMELS.....	3.78	4	0.946	0.177	IGHV1-69 (0.95)	
.....SENTAVY....	3.60	4	0.900	0.171	IGHV1-69 (0.90)	
..RLVQSGAEVR.....	3.34	4	0.834	0.101	IGHV1-69 (0.83)	
.....ENTAVYYCAR	3.30	4	0.824	0.051	IGHV1-69 (0.82)	
.....VEDTAVYYCAR	3.25	4	0.813	0.033	IGHV1-69 (0.81)	
.....FGTANY.....	3.00	3	1.000	1.000	IGHV1-69 (1.00)	
.....GTFSNYALSWVR.....	3.00	3	1.000	1.000	IGHV1-69 (1.00)	
.....VTLTADR.....	3.00	3	1.000	0.980	IGHV1-69 (1.00)	
.....TSTAYMELSSLR.....	2.98	3	0.993	0.491	IGHV1-69 (0.99)	

Figure 10 - Excerpt of expanded HTML report for file JW16_M03+M03a_HAGG_RF_KM_75kd. The top 30 assigned peptides are shown.

- IgFamily v0.12.3 pseudocode

The pseudocode for the IgFamily program follows. The complete source code can be found at -

https://github.com/Lukah0173/IgFamily

1. Initialise default settings and prompt user to confirm default settings or select custom settings:

   1a. Perform display_menu().

2. If selected, perform msconvert file conversion:

   2a. If (perform_wiff_fileconversion) perform wiff_fileconversion().

3. If selected, perform NOVOR de novo peptide assignment:

   3a. If (perform_novor_denovo) perform novor_denovo().

4. Parse FASTA files into raw data structures:

   4a. For (selected_FASTA_file) perform parse_FASTA();

5. Parse data files into raw data structures:

   5a. If (peptide_assignment_method == PEAKS_database) perform parse_PEAKS_database_peptides().

   5b. If (peptide_assignment_method == PEAKS_denovo) perform parse_PEAKS_denovo_peptides().

   5c. If (peptide_assignment_method == NOVOR_denovo) perform parse_NOVOR_denovo_peptides().

6. Assign raw data structures to designed data structures:

   6a. Perform create_v_peptide_data().

   6b. Perform create_v_peptide_analysis().

   6c. Perform create_v_protein_data().

7. From assigned data structures create BLAST input file and BLAST database:

   7a. Perform create_blastinput().

   7b. Perform create_blastp_database().

8. Create a system process and direct blastp.exe to created input file and database:

   8a. Perform systemcall_blastp().

9. Parse BLAST output file to raw data structures:

   9a Perform parse_homology_data().

10. Transform homology data and associate homology data to peptide and protein data structures:

10a. Perform transform_homology_data().

10b. Perform associate_homology_data_to_peptide_data().

10c. Perform associate_homology_data_to_protein_data().

11. Through homology data association to peptide and protein data determine peptide parameters:

11a. Perform determine_homology_data_parameters().

12. Create protein_analysis data structures:

12a. Perform create_v_protein_analysis().

13. Determine protein_analysis parameters and sort for highest multinomial density proteins:

13a. Perform determine_protein_analysis_parameters().

13b. Perform sort_v_protein_analysis().

14. Determine most likely germline allele representation and create new BLAST input:

14a. Perform select_protein_analysis_by_score().

14b. Perform create_blastp_database_refined().

15. Create a system process and direct blastp.exe to created input file and database:

15a. Perform systemcall_blastp().

16. Parse blastp output file to raw data structures:

16a Perform parse_homology_data().

17. Transform homology data and associate homology data to peptide and protein data structures:

17a. Perform transform_homology_data().

17b. Perform associate_homology_data_to_peptide_data().

17c. Perform associate_homology_data_to_protein_data().

18. Align query data to subject data:

18. create_blastp_query_alignment().

19. Through homology data association to peptide and protein data, determine homology_density and score:

19a. Perform determine_homology_data_parameters().

20. Create protein_analysis data structure:

20a. Perform create_v_protein_analysis().

21. Conjugate homology_data and protein_analysis score through iterative process, until cluster condition is achieved:

21a. While (count_cluster_proportion() > select_n_gene_families) Perform conjugate_homology ().

22. Determine protein_analysis parameters and sort for highest multinomial density proteins:

22a. Perform determine_protein_score_density().

22b. Perform sort_v_protein_analysis().

23. Create multinomial data frame:

23a. Perform create_multinomial_data().

24. Create report and output data:

24a. Perform fout_v_peptide_data().

24b. Perform fout_v_protein_data().

24c. Perform fout_v_peptide_analysis().

24d. Perform fout_v_protein_analysis().

24e. Perform fout_v_homology_data().

24f. Perform fout_multinomial().

24g. Perform fout_multinomial_element().

24h. Perform fout_multinomial_element_no_match().

24i. Perform fout_multinomial_contaminant_report().

24j. Perform fout_multinomial_contaminants_list().

24k. Perform fout_multinomial_protein_score().

24l. Perform fout_multinomial_protein_density().

24m. Perform fout_HTML_report().

24n. Perform fout_filesystem_data().

```
-- IgFamily v0.12.3 --


Current settings:

FASTA file -                    IGHV_IGLV_IGKV_CONT_UNIPROT_20160827.fasta
Spectra assignment method -     PEAKS de novo

Select FASTA file:                          [F]
Select spectra assignment method:           [P]
Continue with current settings:             [X]

--> X



reading root directory...

* root_directory\Will_Murray-Brown\20160927\WM16_B02+B02a_HAGG_isolated_RF_MJ2\

* reading: WM16_B02+B02a_HAGG_isolated_RF_MJ2   version - v0.12.2g


* parsing FASTA file... FASTA\IGHV_IGLV_IGKV_CONT_UNIPROT_20160827.fasta

FASTA accessions parsed - 20795



parsing data for file WM16_B02+B02a_HAGG_isolated_RF_MJ2...

--- PEAKS de novo peptides file found


* parsing WM16_B02+B02a_HAGG_isolated_RF_MJ2 PEAKS de novo peptides...

peptides parsed - 7431



creating data structures for file WM16_B02+B02a_HAGG_isolated_RF_MJ2...
...data structures assigned


analysing homology for file WM16_B02+B02a_HAGG_isolated_RF_MJ2...

* * * Calling blastp.exe * * *



Building a new DB, current time: 11/13/2016 14:34:49
New DB name:   C:\Users\dyke0\IgFamily\blast_directory\FPF_blastpdb
New DB title:  blastp_database.fasta
Sequence type: Protein
Deleted existing Protein BLAST database named C:\Users\dyke0\IgFamily\blast_directory\FPF_blastpdb
Keep Linkouts: T
Keep MBits: T
Maximum file size: 1000000000B
Adding sequences from FASTA; added 20429 sequences in 0.568598 seconds.



* * * Closing blastp.exe * * *

...homology analysis complete


creating homology data structures for file WM16_B02+B02a_HAGG_isolated_RF_MJ2...


scoring proteins...
...proteins scored
```

```
determining most-probable germline representation...


reanalysing homology for file WM16_B02+B02a_HAGG_isolated_RF_MJ2...

* * * Calling blastp.exe * * *



Building a new DB, current time: 11/13/2016 14:34:52
New DB name:   C:\Users\dyke0\IgFamily\blast_directory\FPF_blastpdb
New DB title:  blastp_database.fasta
Sequence type: Protein
Deleted existing Protein BLAST database named C:\Users\dyke0\IgFamily\blast_directory\FPF_blastpdb
Keep Linkouts: T
Keep MBits: T
Maximum file size: 1000000000B
Adding sequences from FASTA; added 180 sequences in 0.00522475 seconds.



* * * Closing blastp.exe * * *

...homology analysis complete


creating homology data structures for file WM16_B02+B02a_HAGG_isolated_RF_MJ2...


scoring proteins...
...proteins scored

training protein scores...

...iteration 1 with 24 gene families
...iteration 34 with 7 gene families


creating multinomial data frames...



producing summary reports...
...outputting data reports for WM16_B02+B02a_HAGG_isolated_RF_MJ2
...outputting homology data report for WM16_B02+B02a_HAGG_isolated_RF_MJ2
...outputting multinomial reports for WM16_B02+B02a_HAGG_isolated_RF_MJ2
...outputting multinomial data frame for WM16_B02+B02a_HAGG_isolated_RF_MJ2
...outputting multinomial peptide list for WM16_B02+B02a_HAGG_isolated_RF_MJ2
...outputting filtered multinomial peptide list for WM16_B02+B02a_HAGG_isolated_RF_MJ2
...outputting contaminants report for WM16_B02+B02a_HAGG_isolated_RF_MJ2
...outputting contaminants list for WM16_B02+B02a_HAGG_isolated_RF_MJ2
...outputting protein score comparison for WM16_B02+B02a_HAGG_isolated_RF_MJ2
...outputting html report for WM16_B02+B02a_HAGG_isolated_RF_MJ2




program complete...


input any key to exit...

>
```

Consideration

To clarify the use of the IgFamily program and an example is considered. The model is described with detail in the Story and Function chapters. Reference terms for this chapter are noted in Table 8.

- Example: B02+B02a - Sjögren's Syndrome serum with HAGG purification

The patient has primary Sjögren's Syndrome and serum has been collected. The sample has been collected at a stage of high disease activity shortly before a treatment regime was initiated. The serum was purified with the HAGG purification method and prepared as a technical duplicate. The gene family summary report for the top five ranking gene families is shown in Figure 11. The expanded reports for the top three ranking gene families are shown in Figure 12, Figure 13, and Figure 14.

Table 8 - Terms associated with the IgFamily program.

Term	Definition
$\boldsymbol{p}$	The peptide $\boldsymbol{p}$ of all peptides $\boldsymbol{p} \in \boldsymbol{P}$.
$\boldsymbol{q}$	The protein sequence aligned to by peptide $\boldsymbol{p}$.
$\boldsymbol{g}$	The protein $\boldsymbol{g}$ of all peptides $\boldsymbol{g} \in \boldsymbol{G}$.
$\#\boldsymbol{p}$	The number of observed spectra assigned as the peptide $\boldsymbol{p}$.
$\boldsymbol{\gamma}$	The naivety of the model. This value reduces the weighting of the gene family evidence such that the homology of peptides are more like the values initially assigned by the BLAST program.
$\boldsymbol{E_g(p)}$	The sequence mismatch count of the peptide $\boldsymbol{p}$ for the protein $\boldsymbol{g}$. Each amino acid mismatch of the peptide sequence to the database sequence increases $\boldsymbol{E_g(p)}$ by a value of 1. The value of $\boldsymbol{E_g(p)}$ has a role in defining the rate of amino acid substitution.
$\boldsymbol{D_g(p)}$	The homology density of the peptide $\boldsymbol{p}$ for the protein $\boldsymbol{g}$. It represents the weight of the homology of the peptide $\boldsymbol{p}$ in for a particular gene family $\boldsymbol{g}$ in comparison to all other proteins it could be assigned to. For example a $\boldsymbol{D_g(p)}$ of 0.400 states that the peptide $\boldsymbol{p}$ has a 40% likelihood of originating from the protein $\boldsymbol{g}$.
$\boldsymbol{C_g(p)}$	The conjugated homology density of the peptide $\boldsymbol{p}$ for the protein $\boldsymbol{g}$. It represents the weight of the conjugated homology of the peptide $\boldsymbol{p}$ in for a particular gene family $\boldsymbol{g}$ in comparison to all other proteins it could be assigned to. For example a $\boldsymbol{C_g(p)}$ of 0.700 states that the peptide $\boldsymbol{p}$ has a 70% likelihood of originating from the protein $\boldsymbol{g}$.
$\boldsymbol{p(g)}$	The multinomial density of the gene family $\boldsymbol{g}$. This represents the proportion of the sample that is gene family $\boldsymbol{g}$. For example a $\boldsymbol{p(g)}$ of 0.200 states that the gene family $\boldsymbol{g}$ is 20% of the gene family proteins in the sample. Note that while the model considers all proteins in the database, only the gene family proteins are considered for the mulitnomial density $\boldsymbol{p(g)}$.

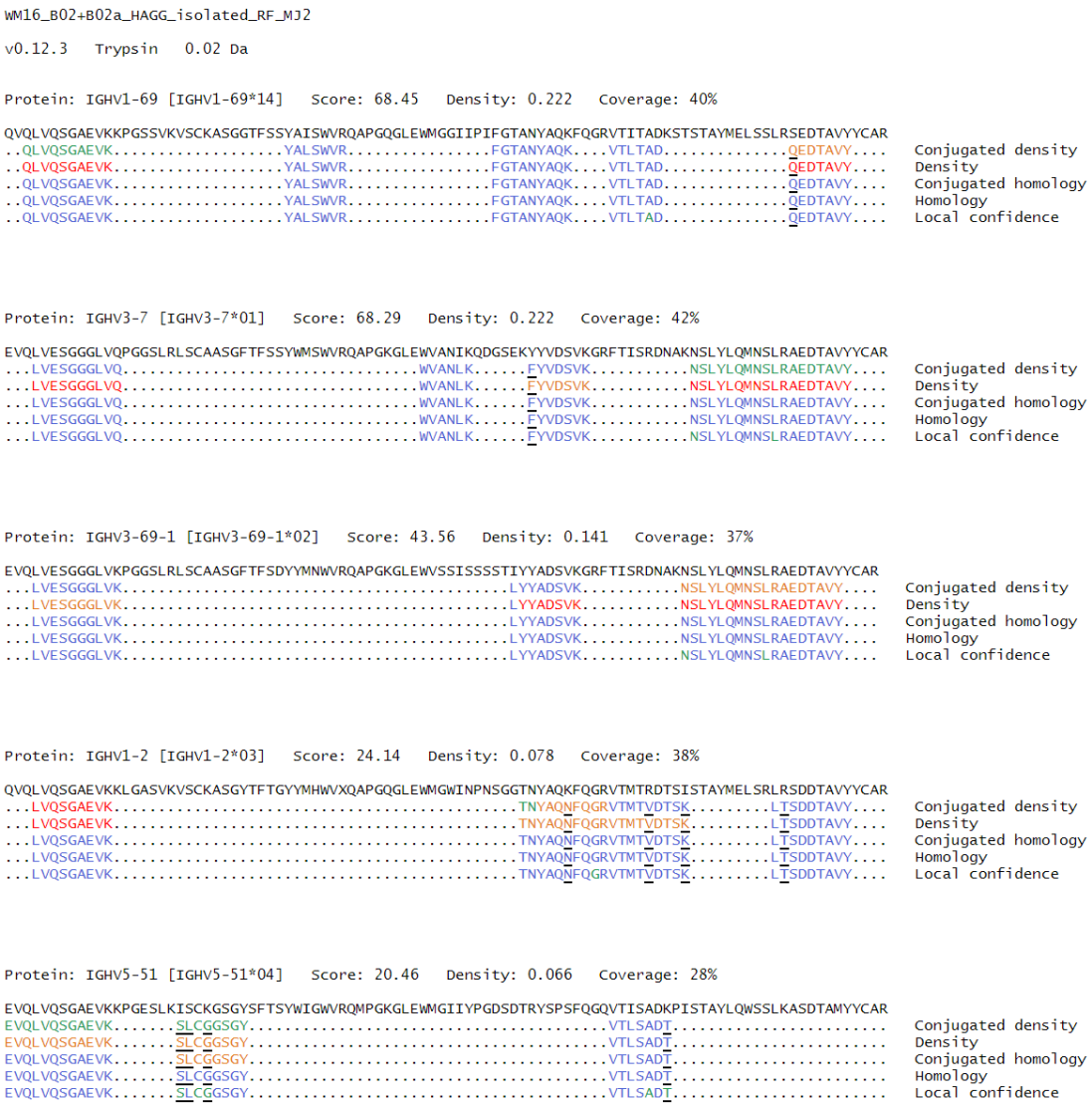


Figure 11 - Gene family summary for file WM16_B02+B02a_HAGG_isolated_RF_MJ2.

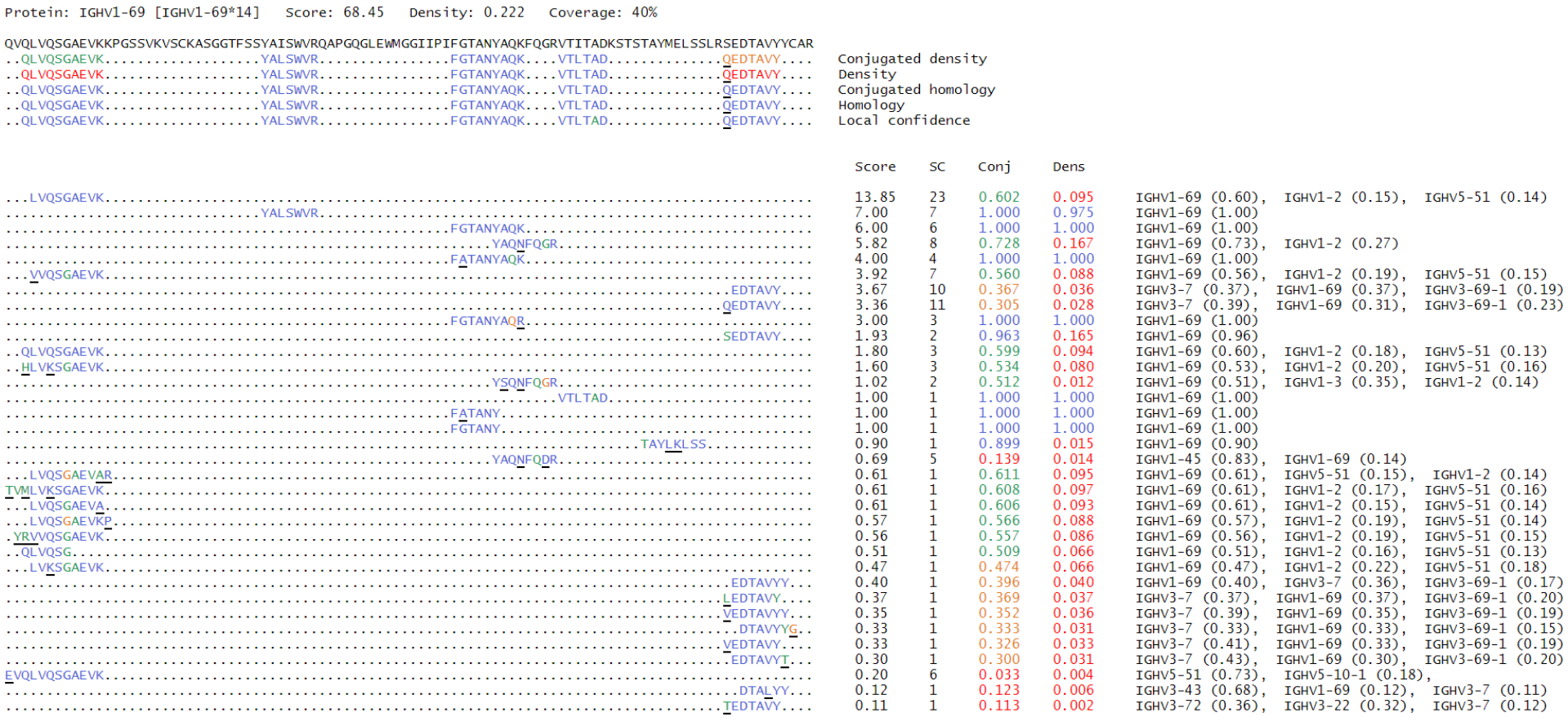


Figure 12 - Gene family #1 from file WM16_B02+B02a_HAGG_isolated_RF_MJ2.

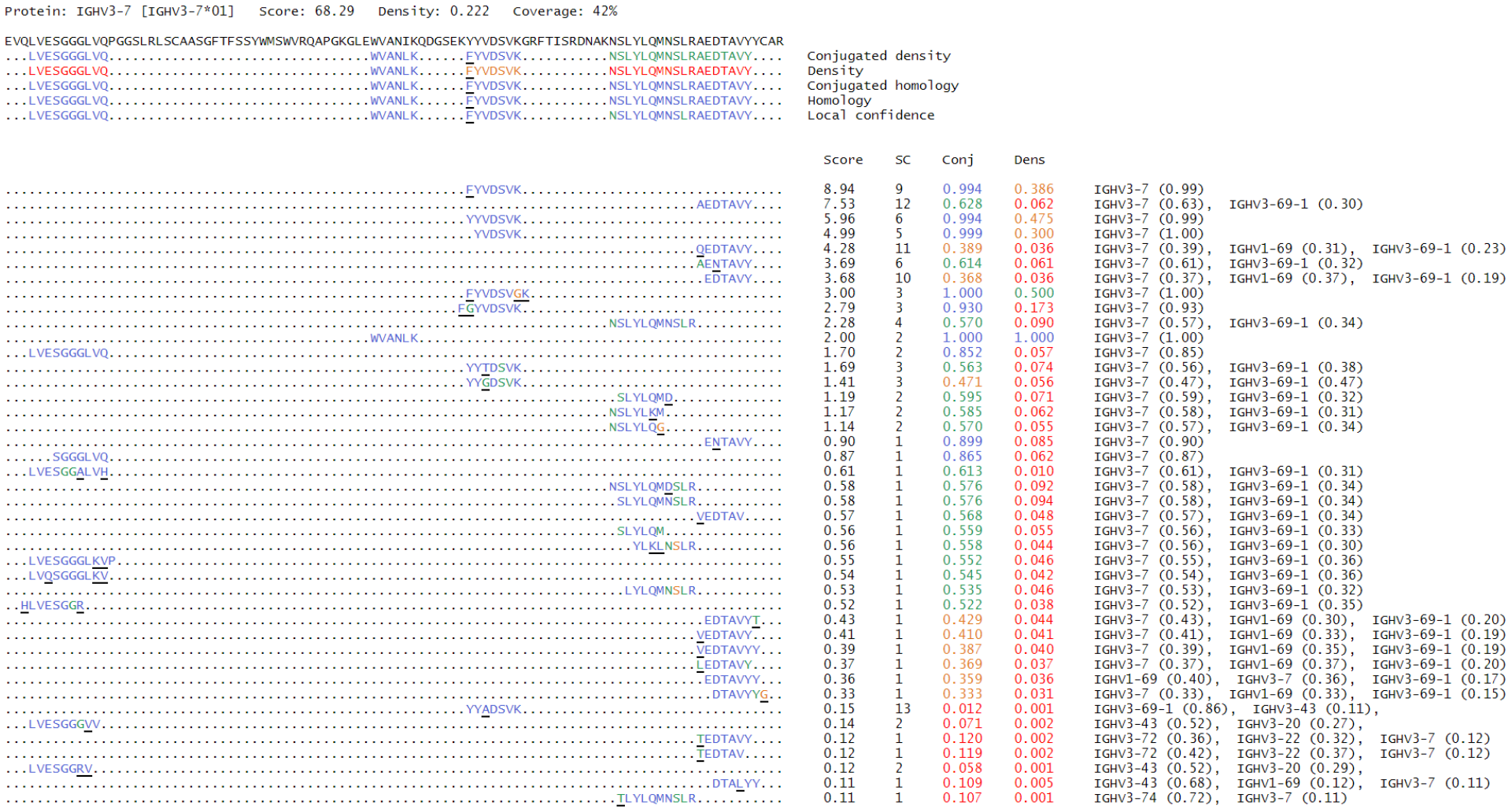


Figure 13 - Gene family #2 from file WM16_B02+B02a_HAGG_isolated_RF_MJ2.

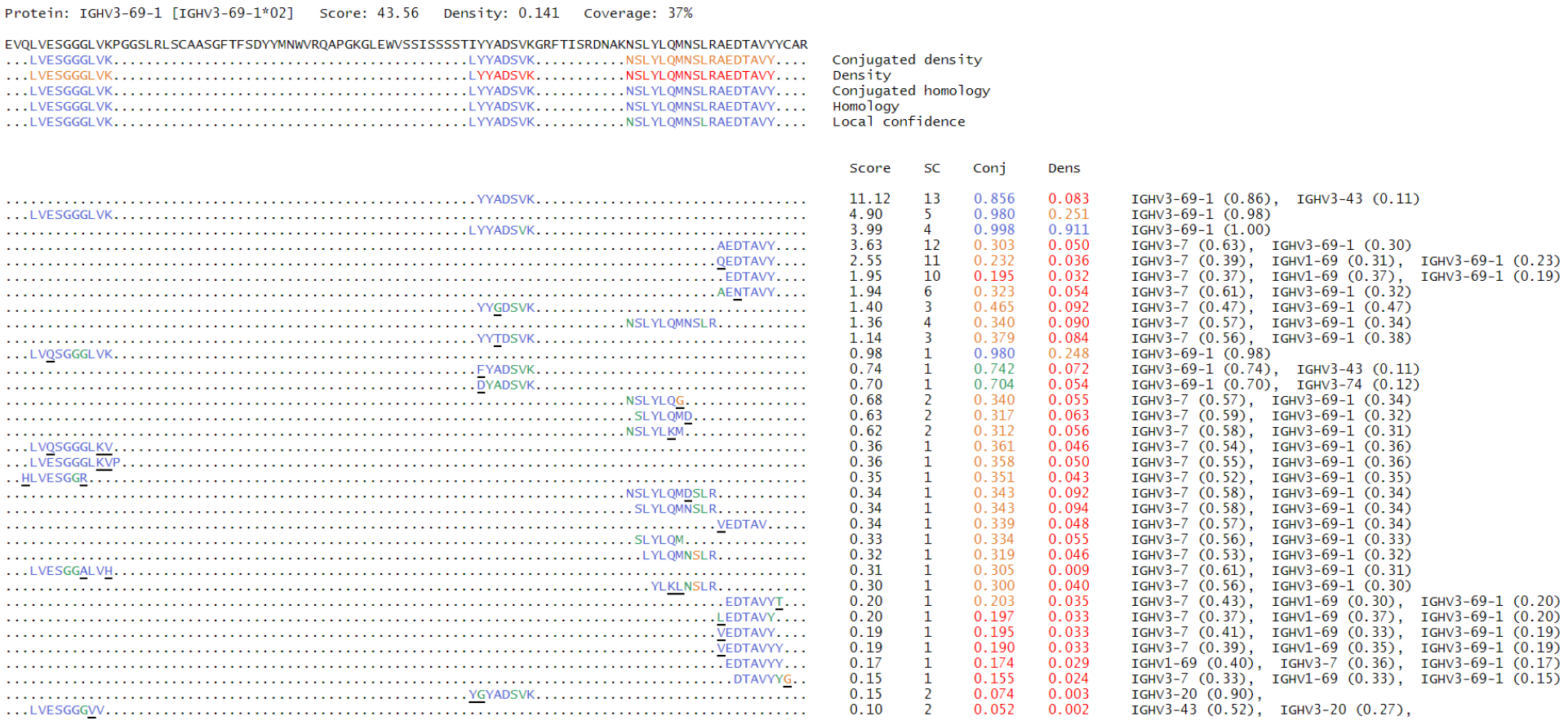
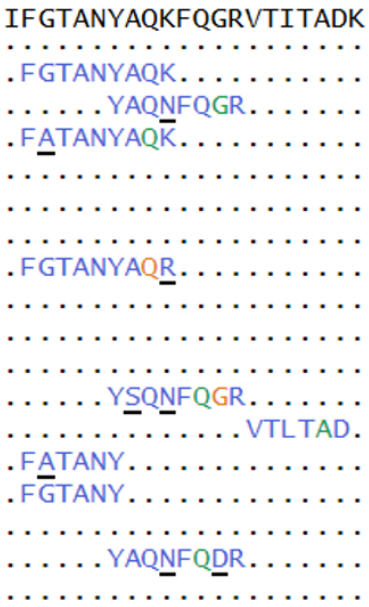


Figure 14 - Gene family #3 from file WM16_B02+B02a_HAGG_isolated_RF_MJ2.

- Where do peptides belong?

In the B02+B02a example the top ranked gene family is IGHV1-69 with a multinomial density $\boldsymbol{p(IGHV1\text{-}69)}$ of 0.222. Figure 15 shows a grouping of distinct peptides that support the IGHV1-69 gene family. These peptides are spatially leading from the CDR2 at the FGTA motif to the FR3. In particular, the peptide grouping FGTANYAQK, FATANYAQK, FGTANYAQR, FATANY, and FGTANY and the peptide VTLTAD are strongly distinct for IGHV1-69 with a $\boldsymbol{D_{IGHV1\text{-}69}(p)}$ of 0.999 with little change to $\boldsymbol{C_{IGHV1\text{-}69}(p)}$ (there is an increase of $\boldsymbol{C_{IGHV1\text{-}69}(p)}$ beyond the decimal precision). Table 9 lists other potential gene family assignments above the threshold value. The peptides FATANYAQK and FGTANYAQR have an almost certain $\boldsymbol{D_{IGHV1\text{-}69}(p)}$ and $\boldsymbol{C_{IGHV1\text{-}69}(p)}$ considering each has an amino acid substitution from the IGHV1-69 germline. In addition, the peptide YAISWVR is distinct for IGHV1-69 with an increase of $\boldsymbol{D_{IGHV1\text{-}69}(YALSWVR)}$ of 0.975 to a $\boldsymbol{C_{IGHV1\text{-}69}(YALSWVR)}$ of 0.999 (Table 10, not shown in Figure 15). The peptide grouping YAQNFQGR, YSQNFQGR, and YAQNFQDR were observed with 8, 2, and 5 spectra suggesting a relatively high sample abundance. YAQNFQGR was initially shared among IGHV1 and IGHV7 gene families and the overall evidence in the sample allowed a stronger assignment to IGHV1-69 with a $\boldsymbol{D_{IGHV1\text{-}69}(YAQNFQGR)}$ increasing from 0.167 to a $\boldsymbol{C_{IGHV1\text{-}69}(YAQNFQGR)}$ of 0.729 (Table 11). Such a determination is a combination of robust evidence for IGHV1-69, low evidence for competing assignments, the power of the clustering rate, and the naivety of the model $\boldsymbol{\gamma}$. Distinct peptides for IGHV1-69 are centred in a small region and of these YAQNFQGR is the most abundant. It is through conjugation that the role of this peptide in supporting IGHV1-69 is revealed. The peptides YSQNFQGR and YAQNFQDR are also conjugated as a reasonable association to IGHV1-69. However these peptides have both a higher mismatch count of 2 and have a much lower homology density $\boldsymbol{D_{IGHV1\text{-}69}(p)}$. Although there is still a substantial increase with $\boldsymbol{C_{IGHV1\text{-}69}(p)}$, it is difficult to assign YSQNFQGR and YAQNFQDR to IGHV1-69 with certainty - at least some of the evidence is shared with the competing assignments, and the naivety of the model also influences restraint.



$\boldsymbol{p}$	$\#\boldsymbol{p}$	$\boldsymbol{E_{1-69}(p)}$	$\boldsymbol{D_{1-69}(p)}$	$\boldsymbol{C_{1-69}(p)}$
FGTANYAQK	6	0	0.999	0.999
YAQNFQGR	8	1	0.167	0.729
FATANYAQK	1	1	0.999	0.999
FGTANYAQR	3	1	0.999	0.999
YSQNFQGR	2	2	0.012	0.512
VTLTAD	1	0	0.999	0.999
FATANY	1	1	0.999	0.999
FGTANY	1	0	0.999	0.999
YAQNFQDR	5	2	0.014	0.139

Figure 15 - Peptides from CDR2 into FR3 support gene family IGHV1-69.

Table 9 - The peptides FGTANYAQK, FATANYAQK, and FGTANYAQR are strongly distinct for IGHV1-69. Note that the peptides FATANYAQK and FGTANYAQR have a substitution from the IGHV1-69 germline.

p	q	g	$E_g(p)$	$D_g(p)$	$C_g(p)$
FGTANYAQK	FGTANYAQK	IGHV1-69	0	0.99999	0.99999
FGTANYAQK	GGTNYAQK	IGHV1-2	3	4.99E-07	1.42E-07
FGTANYAQK	GNTNYAQK	IGHV1-45	3	5.62E-07	4.53E-08
FGTANYAQK	GNTNYAQK	IGHV1-18	3	5.62E-07	2.56E-73
FGTANYAQK	GNTNYAQK	IGHV1-58	3	4.99E-07	2.19E-90

p	q	g	$E_g(p)$	$D_g(p)$	$C_g(p)$
FATANYAQK	FGTANYAQK	IGHV1-69	1	0.99999	0.99999
FATANYAQK	NYAQK	IGHV1-45	4	6.17E-07	4.98E-08

p	q	g	$E_g(p)$	$D_g(p)$	$C_g(p)$
FGTANYAQR	FGTANYAQK	IGHV1-69	1	0.99999	0.99999
FGTANYAQR	GGTNYAQK	IGHV1-2	4	6.54E-07	1.86E-07
FGTANYAQR	GNTNYAQK	IGHV1-45	4	8.47E-07	6.85E-08
FGTANYAQR	GNTNYAQK	IGHV1-18	4	8.47E-07	3.87E-73
FGTANYAQR	GNTNYAQK	IGHV1-58	4	6.54E-07	2.88E-90

Table 10 - The peptide YAISWVR supports gene family IGHV1-69.

p	q	g	$E_g(p)$	$D_g(p)$	$C_g(p)$
YAISWVR	YAISWVR	IGHV1-69	0	0.97528	0.99998
YAISWVR	YAMNWVR	IGHV7-4-1	2	0.00012	1.43E-05
YAISWVR	YAMHWVR	IGHV3-43	2	0.00012	9.75E-06
YAISWVR	YAMHWVR	IGHV1-3	2	0.00012	9.80E-07
YAISWVR	YAMSWFR	IGHV3-49	2	0.00012	4.74E-08
YAISWVR	YAMHWVR	IGHV3-30-3	2	0.00012	4.48E-08
YAISWVR	YAMSWVR	IGHV3-23	1	0.01442	1.44E-14
YAISWVR	YAMHWVR	IGHV3-64	2	0.00012	2.50E-18
YAISWVR	YALHWVR	IGHV3-47	1	0.00943	2.88E-169

Table 11 - The peptides YAQNFQGR, YSQNFQGR, and YAQNFQDR are initially distributed among IGHV1 and IGHV7 gene families. Evidence in the sample supports association to IGHV1-69.

p	q	g	$E_g(p)$	$D_g(p)$	$C_g(p)$
YAQNFQGR	YAQKFQGR	IGHV1-69	1	0.16743	0.72765
YAQNFQGR	YAQKFQGR	IGHV1-2	1	0.20337	0.27000
YAQNFQGR	YAQGFTGR	IGHV7-81	2	0.00162	0.00089
YAQNFQGR	YAQGFTGR	IGHV7-4-1	2	0.00128	0.00064
YAQNFQGR	YAQKFQDR	IGHV1-45	2	0.00144	0.00053
YAQNFQGR	YAEKFQGR	IGHV1-69-2	2	0.00273	0.00021
YAQNFQGR	YSQEFQGR	IGHV1-3	2	0.00224	7.65E-05
YAQNFQGR	YAQKFQGR	IGHV1-24	1	0.24452	1.41E-15
YAQNFQGR	YAQKFQGR	IGHV1-46	1	0.16743	3.85E-20
YAQNFQGR	YAQKFQGR	IGHV1-8	1	0.20337	7.80E-24
YAQNFQGR	YAKKFQGR	IGHV1-68	2	0.00162	6.12E-33
YAQNFQGR	YAQKLQGR	IGHV1-18	2	0.00181	3.76E-69
YAQNFQGR	YAQKFQER	IGHV1-58	2	0.00113	2.27E-86

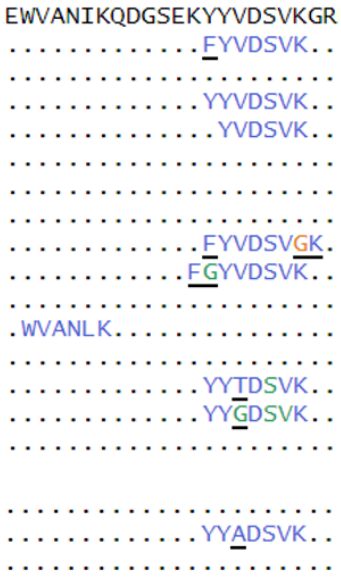
p	q	g	$E_g(p)$	$D_g(p)$	$C_g(p)$
YSQNFQGR	YAQKFQGR	IGHV1-69	2	0.01151	0.51210
YSQNFQGR	YSQEFQGR	IGHV1-3	1	0.94539	0.34562
YSQNFQGR	YAQKFQGR	IGHV1-2	2	0.01040	0.14113
YSQNFQGR	YAQGFTGR	IGHV7-4-1	3	8.81E-05	0.00045
YSQNFQGR	YAQGFTGR	IGHV7-81	3	7.79E-05	0.00044
YSQNFQGR	YAQKFQDR	IGHV1-45	3	6.86E-05	0.00026
YSQNFQGR	YAQKFQGR	IGHV1-24	2	0.01151	6.77E-16
YSQNFQGR	YAQKFQGR	IGHV1-46	2	0.01040	2.45E-20
YSQNFQGR	YAQKFQGR	IGHV1-8	2	0.01040	4.08E-24
YSQNFQGR	YAQKLQGR	IGHV1-18	3	8.81E-05	1.87E-69
YSQNFQGR	YAQKFQER	IGHV1-58	3	6.86E-05	1.41E-86

p	q	g	$E_g(p)$	$D_g(p)$	$C_g(p)$
YAQNFQDR	YAQKFQDR	IGHV1-45	1	0.92785	0.83098
YAQNFQDR	YAQKFQGR	IGHV1-69	2	0.01372	0.13872
YAQNFQDR	YAQKFQGR	IGHV1-2	2	0.00983	0.03028
YAQNFQDR	YAEKFQGR	IGHV1-69-2	3	0.00013	2.31E-05
YAQNFQDR	YSQEFQGR	IGHV1-3	3	8.81E-05	6.98E-06
YAQNFQDR	YAQKFQGR	IGHV1-24	2	0.01372	1.83E-16
YAQNFQDR	YAQKFQGR	IGHV1-46	2	0.01103	5.90E-21
YAQNFQDR	YAQKFQGR	IGHV1-8	2	0.00983	8.75E-25
YAQNFQDR	YAKKFQGR	IGHV1-68	3	8.81E-05	7.73E-34
YAQNFQDR	YAQKFQER	IGHV1-58	2	0.01372	6.56E-85

There are important considerations with the peptides YAQNFQGR, YSQNFQGR, and YAQNFQDR. Notably, even with a view of all possible originating gene family germlines, there must be at least one substitution from the germline sequence. The consequence of this presupposes that a peptide can lose identity to a germline and that assignment requires a holistic approach to the data. This is evident for the YSQNFQGR peptide, where the most conservative association is to IGHV1-3 with a single substitution compared to the two that are required for a IGHV1-69 germline. Indeed, the density $\boldsymbol{D_{IGHV1\text{-}69}(YSQNFQGR)}$ is a paltry 0.012 before conjugation raises it to a $\boldsymbol{C_{IGHV1\text{-}69}(YSQNFQGR)}$ of 0.512.

The association of the peptides YAQNFQGR, YSQNFQGR, and YAQNFQDR to IGHV1-69 opens questions about the rate of amino acid substitution. The peptide YSQNFQGR needs the supporting evidence of more readily assignable peptides, particularly the FGTANYAQK group and YAISWVR, to overcome the conservative association to IGHV1-3. In general there are peptides that are difficult to assign to a gene family through insufficient evidence. Suppose the FGTANYAQK group and YAISWVR were not observed. It would be reasonable that the peptides YAQNFQGR, YSQNFQGR, and YAQNFQDR would be associated to IGHV1-3 and the association would likely be incorrect to the reality of the sample. This is not a pitfall of the model however, it makes its claim from the ground of the available data. The problem becomes one of the resolving power of the data. This is examined in greater detail later in this chapter.

The gene family with the second highest multinomial density is IGHV3-7 with a $\boldsymbol{p(IGHV3\text{-}7)}$ of 0.222. There is a region covering the FR2, CDR2, and FR3 that support IGHV3-7. However, unlike for the distinct peptides for IGHV1-69, the values of $\boldsymbol{D_{IGHV3\text{-}7}(p)}$ suggests that there is at least one other gene family that could explain the observed peptides. Analysis through the model supports assignment of the peptides FYVDSVK, YYVDSVK, and YVDSVK for IGHV3-7 with a respective increase of $\boldsymbol{D_{IGHV3\text{-}7}(p)}$ from 0.386, 0.475, and 0.300 to a $\boldsymbol{C_{IGHV3\text{-}7}(p)}$ of 0.994, 0.994, and 0.999. The weight of the evidence of IGHV3-7 also supports the peptides YYTDSVK, YYGDSVK, and FGYVDSVK, although there remains a significant uncertainty. This uncertainty is a consequence of a competing gene family, IGHV3-69-1, that will be described shortly.



p	$\#p$	$E_{3-7}(p)$	$D_{3-7}(p)$	$C_{3-7}(p)$
FYVDSVK	9	1	0.386	0.994
YYVDSVK	6	0	0.475	0.994
YVDSVK	5	0	0.300	0.999
FYVDSVGK	3	3	0.500	0.999
FGYVDSVK	3	2	0.173	0.930
WVANLK	2	0	0.999	0.999
YYTDSVK	3	1	0.074	0.563
YYGDSVK	3	1	0.056	0.471
YYADSVK	13	1	0.001	0.012

Figure 16 - Peptides covering the FR2, CDR2, and FR3 support gene family IGHV3-7.

The third ranked gene family is IGHV3-69-1 with a multinomial density $\boldsymbol{p(IGHV3\text{-}69\text{-}1)}$ of 0.141. Peptides supporting IGHV3-69-1 cover a narrow region at the end of the CDR2 leading into the FR3. The IGHV3-69-1 gene family has scored well considering the small coverage of significant peptides. This is largely a result of the distinct peptides YYADSVK and IYYADSVK. The peptide IYYADSVK was initially distinct with a $\boldsymbol{D_{IGHV3\text{-}69\text{-}1}(LYYADSVK)}$ of 0.911 increasing to a $\boldsymbol{C_{IGHV3\text{-}69\text{-}1}}$ $\boldsymbol{(LYYADSVK)}$ of 0.999. However, the peptide YYADSVK receives a seemingly disproportionate increase of $\boldsymbol{D_{IGHV3\text{-}69\text{-}1}(YYADSVK)}$ from 0.083 to a $\boldsymbol{C_{IGHV3\text{-}69\text{-}1}}$ $\boldsymbol{(YYADSVK)}$ of 0.856. The consequence of this is that the majority of the gene family density $\boldsymbol{p(IGHV3\text{-}69\text{-}1)}$ for IGHV3-69-1 is due to the high spectral count of YYADSVK and its *a posteriori* distinctiveness. This seems reasonable, at least if the peptide IYYADSVK could be known to also originate from IGHV3-69-1. Although each peptide in its own right gives an impression that the peptide IYYADSVK should be distinct for IGHV3-69-1, it turns out that the assignment is complicated by the potential of clonal divergence from an initial clonal type. This theme is important and is considered later in this chapter.

The peptides YYGDSVK and YYTDSVK are not well resolved for IGHV3-69-1 with a $\boldsymbol{C_{IGHV3\text{-}69\text{-}1}(p)}$ of 0.465 and 0.379. There is a competing evidence for a IGHV3-7 assignment, and the peptides could be a product of a IGHV3-7 germline substitution from the peptide YYVDSVK. In contrast the peptides FYADSVK and DYADSVK only require a single substitution from a IGHV3-69-1 germline YYADSVK, but require two substitutions from a IGHV3-7 germline YYVDSVK. Table 12 demonstrates the values of $\boldsymbol{D_g(p)}$ and $\boldsymbol{C_g(p)}$ between the IGHV3-7 and IGHV3-69-1 gene families. The values of $\boldsymbol{C_g(p)}$ suggest that the grouping of peptides is able to be resolved excepting the two peptides YYGDSVK and YYTDSVK. This is not disastrous at least, the majority of peptides are able to be confidently assigned to a gene family and the areas of overlap make sense with the substitution rate implicit in the model. However, it could also be argued that such peptides originated entirely from IGHV3-7 or IGHV3-69-1 and that the rate of substitution is much greater. Alternatively, substitutions may be altogether quite unlikely.

$\boldsymbol{p}$	$\boldsymbol{\#p}$	$\boldsymbol{E_{3-69-1}(p)}$	$\boldsymbol{D_{3-69-1}(p)}$	$\boldsymbol{C_{3-69-1}(p)}$
YYADSVK	13	0	0.083	0.856
LYYADSVK	4	0	0.911	0.998
YYGDSVK	3	1	0.092	0.465
YYTDSVK	3	1	0.084	0.379
FYADSVK	1	1	0.072	0.742
DYADSVK	1	1	0.054	0.704

Figure 17- Peptides covering the CDR2 and FR3 support gene family IGHV3-69-1.

Table 12 - Values of $\boldsymbol{D_g(p)}$ and $\boldsymbol{C_g(p)}$ for the IGHV3-7 and IGHV3-69-1 gene families. The peptides YYGDSVK and YYTDSVK are not able to be resolved to a distinct origin.

$\boldsymbol{p}$	$\#\boldsymbol{p}$	$\boldsymbol{D_{IGHV3\text{-}7}(p)}$	$\boldsymbol{D_{IGHV3\text{-}69\text{-}1}(p)}$	$\boldsymbol{C_{IGHV3\text{-}7}(p)}$	$\boldsymbol{C_{IGHV3\text{-}69\text{-}1}(p)}$
YYVDSVK	6	0.475	0.005	0.994	0.006
FYVDSVK	9	0.386	0.003	0.994	0.005
YVDSVK	5	0.300	0.000	0.998	0.000
YYTDSVK	3	0.074	0.084	0.563	0.379
YYGDSVK	3	0.056	0.092	0.471	0.465
YYADSVK	13	0.001	0.082	0.012	0.856
IYYADSVK	4	0.000	0.911	0.000	1.000
FYADSVK	1	0.001	0.072	0.011	0.742
DYADSVK	1	0.000	0.054	0.000	0.704

- The problem with parameters

Without the aid of known outcomes to best fit the parameters, the selection of suitable values is a complex task, even guided by intuitive results. In general there exists at least one set of parameters that optimally represents that data while also being bound by the constraints of the model. With a full probability model the data could be best fit by maximising a performance measure - such as a maximum entropy definition of clustering. Such a model is proposed in the Future chapter. Currently the model uses only the mean multinomial likelihood value of each assignment and doesn't consider variance. The model is fit by supposing some definite amount of gene family proteins exist in the sample. In general the most likely number of gene family proteins can be found by considering the change in assignment variance as this number is adjusted. At least, the model can be fit subjectively be testing many parameters manually. Looking at the effects of adjusting each of the parameters in turn is a simple way of investigating this. Table 13 displays the default parameters that have been selected. Table 14, Table 15, and Table 16 examine the role of each parameter with varying assumptions.

Table 13 - The parameter values used in the IgFamily program.

Parameter	Value	Definition
$\boldsymbol{\alpha_1}$	3.5	The parameter value for the weight of the homology score. An increase in this value places more precedence on the relative differences in homology.
$\boldsymbol{\alpha_2}$	0.3	The parameter value for the peptide sequence mismatch penalty. A decrease in this value emphasises homology by sequence identity.
$\boldsymbol{\eta}$	7	The quantity of gene families assumed to be in the sample. This value is set to be conservative.
$\boldsymbol{\theta}$	0.80	The cluster proportion value of the multinomial density. The IgFamily program regresses a model fit until $\boldsymbol{\theta}$ of the multinomial density is contained in the top $\boldsymbol{\eta}$ gene families.

Table 14 - Effect of adjusting parameter α_1.

$\boldsymbol{p}$	$\boldsymbol{q}$	$\boldsymbol{g}$	$\boldsymbol{E_g(p)}$	$\boldsymbol{D_g(p)}$ ∀! $\boldsymbol{\alpha_1 = 2.5}$	$\boldsymbol{D_g(p)}$ ∀! $\boldsymbol{\alpha_1 = 3.5}$	$\boldsymbol{D_g(p)}$ ∀! $\boldsymbol{\alpha_1 = 4.5}$
FGTANYAQK	FGTANYAQK	IGHV1-69	0	0.99987	0.99999	0.99999
FGTANYAQK	GNTNYAQK	IGHV1-18	3	3.43E-05	5.62E-07	9.19E-09
FGTANYAQK	GNTNYAQK	IGHV1-45	3	3.43E-05	5.62E-07	9.19E-09
FGTANYAQK	GGTNYAQK	IGHV1-2	3	3.15E-05	4.99E-07	7.89E-09
FGTANYAQK	GNTNYAQK	IGHV1-58	3	3.15E-05	4.99E-07	7.89E-09

$\boldsymbol{p}$	$\boldsymbol{q}$	$\boldsymbol{g}$	$\boldsymbol{E_g(p)}$	$\boldsymbol{D_g(p)}$ ∀! $\boldsymbol{\alpha_1 = 2.5}$	$\boldsymbol{D_g(p)}$ ∀! $\boldsymbol{\alpha_1 = 3.5}$	$\boldsymbol{D_g(p)}$ ∀! $\boldsymbol{\alpha_1 = 4.5}$
YAISWVR	YAISWVR	IGHV1-69	0	0.91136	0.97527	0.99298
YAISWVR	YAMSWVR	IGHV3-23	1	0.04493	0.01442	0.00441
YAISWVR	YALHWVR	IGHV3-47	1	0.03317	0.00943	0.00255
YAISWVR	YAMHWVR	IGHV1-3	2	0.00151	0.00012	9.77E-06
YAISWVR	YAMHWVR	IGHV3-9	2	0.00151	0.00012	9.77E-06
YAISWVR	YAMHWVR	IGHV3-30-3	2	0.00151	0.00012	9.77E-06
YAISWVR	YAMHWVR	IGHV3-43	2	0.00151	0.00012	9.77E-06
YAISWVR	YAMSWFR	IGHV3-49	2	0.00151	0.00012	9.77E-06
YAISWVR	YAMHWVR	IGHV3-64	2	0.00151	0.00012	9.77E-06
YAISWVR	YAMNWVR	IGHV7-4-1	2	0.00151	0.00012	9.77E-06

$\boldsymbol{p}$	$\boldsymbol{q}$	$\boldsymbol{g}$	$\boldsymbol{E_g(p)}$	$\boldsymbol{D_g(p)}$ ∀! $\boldsymbol{\alpha_1 = 2.5}$	$\boldsymbol{D_g(p)}$ ∀! $\boldsymbol{\alpha_1 = 3.5}$	$\boldsymbol{D_g(p)}$ ∀! $\boldsymbol{\alpha_1 = 4.5}$
YYVDSVK	YYVDSVK	IGHV3-52	0	0.41602	0.47470	0.49302
YYVDSVK	YYVDSVK	IGHV3-7	0	0.41602	0.47470	0.49302
YYVDSVK	HYVDSVK	IGHV3-16	1	0.01762	0.00568	0.00167
YYVDSVK	YYADSVK	IGHV3-23	1	0.01500	0.00453	0.00125
YYVDSVK	YYADSVK	IGHV3-30	1	0.01500	0.00453	0.00125
YYVDSVK	YYADSVK	IGHV3-30-3	1	0.01500	0.00453	0.00125
YYVDSVK	YYADSVK	IGHV3-30-5	1	0.01500	0.00453	0.00125
YYVDSVK	YYADSVK	IGHV3-33	1	0.01500	0.00453	0.00125
YYVDSVK	YYADSVK	IGHV3-43	1	0.01500	0.00453	0.00125
YYVDSVK	YYADSVK	IGHV3-53	1	0.01500	0.00453	0.00125
YYVDSVK	YYADSVK	IGHV3-66	1	0.01500	0.00453	0.00125
YYVDSVK	YYADSVK	IGHV3-69-1	1	0.01500	0.00453	0.00125
YYVDSVK	YYADSVK	IGHV3-64	1	0.01378	0.00402	0.00107
YYVDSVK	HYADSVK	IGHV3-35	2	0.00062	5.26E-05	4.05E-06
YYVDSVK	YADSVK	IGHV3-20	2	0.00050	3.87E-05	2.73E-06
YYVDSVK	YYADSV	IGHV3-47	2	0.00046	3.39E-05	2.30E-06

Table 15 - Effect of adjusting parameter α_2.

p	q	g	$E_g(p)$	$D_g(p)$ ∀! $\alpha_2 = 0.1$	$D_g(p)$ ∀! $\alpha_2 = 0.3$	$D_g(p)$ ∀! $\alpha_2 = 0.5$
FGTANYAQK	FGTANYAQK	IGHV1-69	0	0.99999	0.99999	0.99955
FGTANYAQK	GNTNYAQK	IGHV1-18	3	5.49E-12	5.62E-07	0.00012
FGTANYAQK	GNTNYAQK	IGHV1-45	3	5.49E-12	5.62E-07	0.00012
FGTANYAQK	GGTNYAQK	IGHV1-2	3	4.88E-12	4.99E-07	0.00011
FGTANYAQK	GNTNYAQK	IGHV1-58	3	4.88E-12	4.99E-07	0.00011

p	q	g	$E_g(p)$	$D_g(p)$ ∀! $\alpha_2 = 0.1$	$D_g(p)$ ∀! $\alpha_2 = 0.3$	$D_g(p)$ ∀! $\alpha_2 = 0.5$
YAISWVR	YAISWVR	IGHV1-69	0	0.99948	0.97527	0.84885
YAISWVR	YAMSWVR	IGHV3-23	1	0.00032	0.01442	0.07503
YAISWVR	YALHWVR	IGHV3-47	1	0.00021	0.00943	0.04906
YAISWVR	YAMHWVR	IGHV1-3	2	5.83E-08	0.00012	0.00387
YAISWVR	YAMHWVR	IGHV3-9	2	5.83E-08	0.00012	0.00387
YAISWVR	YAMHWVR	IGHV3-30-3	2	5.83E-08	0.00012	0.00387
YAISWVR	YAMHWVR	IGHV3-43	2	5.83E-08	0.00012	0.00387
YAISWVR	YAMSWFR	IGHV3-49	2	5.83E-08	0.00012	0.00387
YAISWVR	YAMHWVR	IGHV3-64	2	5.83E-08	0.00012	0.00387
YAISWVR	YAMNWVR	IGHV7-4-1	2	5.83E-08	0.00012	0.00387

p	q	g	$E_g(p)$	$D_g(p)$ ∀! $\alpha_2 = 0.1$	$D_g(p)$ ∀! $\alpha_2 = 0.3$	$D_g(p)$ ∀! $\alpha_2 = 0.5$
YYVDSVK	YYVDSVK	IGHV3-52	0	0.49943	0.47470	0.37809
YYVDSVK	YYVDSVK	IGHV3-7	0	0.49943	0.47470	0.37809
YYVDSVK	HYVDSVK	IGHV3-16	1	0.00013	0.00568	0.02703
YYVDSVK	YYADSVK	IGHV3-23	1	0.00010	0.00453	0.02156
YYVDSVK	YYADSVK	IGHV3-30	1	0.00010	0.00453	0.02156
YYVDSVK	YYADSVK	IGHV3-30-3	1	0.00010	0.00453	0.02156
YYVDSVK	YYADSVK	IGHV3-30-5	1	0.00010	0.00453	0.02156
YYVDSVK	YYADSVK	IGHV3-33	1	0.00010	0.00453	0.02156
YYVDSVK	YYADSVK	IGHV3-43	1	0.00010	0.00453	0.02156
YYVDSVK	YYADSVK	IGHV3-53	1	0.00010	0.00453	0.02156
YYVDSVK	YYADSVK	IGHV3-66	1	0.00010	0.00453	0.02156
YYVDSVK	YYADSVK	IGHV3-69-1	1	0.00010	0.00453	0.02156
YYVDSVK	YYADSVK	IGHV3-64	1	9.05E-05	0.00402	0.01915
YYVDSVK	HYADSVK	IGHV3-35	2	2.53E-08	5.26E-05	0.00150
YYVDSVK	YADSVK	IGHV3-20	2	1.86E-08	3.87E-05	0.00110
YYVDSVK	YYADSV	IGHV3-47	2	1.63E-08	3.39E-05	0.00097

Table 16 - Effect of adjusting parameter $\boldsymbol{\theta}$ for the file WM16_B02+B02a_HAGG_isolated_RF_MJ2.

$\boldsymbol{g}$	$\boldsymbol{p(g)}$ ∀! $\boldsymbol{\theta = 0.75}$	$\boldsymbol{p(g)}$ ∀! $\boldsymbol{\theta = 0.80}$	$\boldsymbol{p(g)}$ ∀! $\boldsymbol{\theta = 0.85}$
IGHV1-69	0.216	0.222	0.256
IGHV3-7	0.206	0.222	0.236
IGHV3-69-1	0.127	0.141	0.141
IGHV1-2	0.080	0.078	0.069
IGHV5-51	0.054	0.066	0.078

Table 17 - Effect of adjusting parameter $\boldsymbol{\eta}$ for the file WM16_B02+B02a_HAGG_isolated_RF_MJ2.

$\boldsymbol{g}$	$\boldsymbol{p(g)}$ ∀! $\boldsymbol{\eta = 5}$	$\boldsymbol{p(g)}$ ∀! $\boldsymbol{\eta = 7}$	$\boldsymbol{p(g)}$ ∀! $\boldsymbol{\eta = 9}$
IGHV1-69	0.276	0.222	0.203
IGHV3-7	0.243	0.222	0.214
IGHV3-69-1	0.140	0.141	0.121
IGHV1-2	0.065	0.078	0.079
IGHV5-51	0.078	0.066	0.053

The more initially distinct a peptide is for a gene family association the less influence there is with a change of parameter. This makes sense: a peptide that is distinctly representative of a protein will be less amenable to variance of the model under a greater body of conditions. In contrast peptides that are heavily shared have in general a smaller proportion of protein association evidence and are more sensitive to the model. It may seem reasonable to focus the inference of the data to those peptides that are most likely to be robust with variance. Indeed, strongly distinct peptides support their associated proteins through a wide distribution of model parameters and are not as likely to succumb to spurious subjectivity. On the other hand, there is good reason to make use of the full body of data, if not only because there is much more of it. More importantly, it is critical to determine when a shared peptide acquires a substitution that gives it a distinct façade.

- A distinct peptide by any other name

Distinct peptides can provide powerful evidence towards a protein of origin. Although the method utilised by the IgFamily program concerns the relative density of the homology, the canonical idea of a distinct peptide often involves determining peptides that uniquely have aligned sequence identity with a protein counterpart. Such a peptide is referred to as a *unique peptide*. Although unique peptides have been fundamental in protein mass spectrometry studies, the sequence similarity among the immunoglobulin gene families brings new challenges.

Consider the peptide FGTANYAQK shown in Table 18. This is a distinct and unique peptide for gene family IGHV1-69 and the model supports this with a $\boldsymbol{D_{IGHV1\text{-}69}(FGTANYAQK)}$ of 0.999. It would be unlikely that another germline protein could acquire the three substitutions necessary to be a IGHV1-69 doppelgänger - not only would it need substantial nucleotide mutation, they would also need to be the right combination of mutations. Table 19 demonstrates another unique peptide for IGHV1-69, YAISWVR. Here there is a small but significant uncertainty with a $\boldsymbol{D_{IGHV1\text{-}69}(YAISWVR)}$ of 0.975. The uncertainty is a result of a possible substitution, most notably from IGHV3-23.

Table 18 - The IGHV1-69 unique peptide FGTANYAQK.

p	q	g	$E_g(p)$	$D_g(p)$
FGTANYAQK	FGTANYAQK	IGHV1-69	0	0.99999
FGTANYAQK	GNTNYAQK	IGHV1-18	3	5.94E-08
FGTANYAQK	GNTNYAQK	IGHV1-45	3	5.94E-08
FGTANYAQK	GGTNYAQK	IGHV1-2	3	5.27E-08
FGTANYAQK	GNTNYAQK	IGHV1-58	3	5.27E-08

Table 19 - The IGHV1-69 unique peptide YAISWVR.

$\boldsymbol{p}$	$\boldsymbol{q}$	$\boldsymbol{g}$	$\boldsymbol{E_g(p)}$	$\boldsymbol{D_g(p)}$
YAISWVR	YAISWVR	IGHV1-69	0	0.97526
YAISWVR	YAMSWVR	IGHV3-23	1	0.01442
YAISWVR	YAMNWVR	IGHV7-4-1	2	0.00012
YAISWVR	YAMHWVR	IGHV3-43	2	0.00012
YAISWVR	YAMHWVR	IGHV1-3	2	0.00012
YAISWVR	YAMHWVR	IGHV3-30-3	2	0.00012
YAISWVR	YAMSWFR	IGHV3-49	2	0.00012
YAISWVR	YAMHWVR	IGHV3-9	2	0.00012
YAISWVR	YAMHWVR	IGHV3-64	2	0.00012

It could be stated that even a 2.5% likelihood of substitution from another gene family is a generous allowance. There is an interesting aspect to this problem that is a consequence of the immunological nature of the gene families. The BLOSUM62 substitution matrix (Figure 2) used through BLAST for the initial homology scoring is created from the phylogenetic divergence of transmembrane proteins. The resulting frequency matrix is both a measure of the physical likelihood of an amino acid exchanging with another and the likelihood of the protein retaining its function with the substitution. For an immunoglobulin this is confounded by the varied nature of the nucleotide mutation rate and the selected proliferation of those immunoglobulins that are successful. The phylogenetic ancestry of the gene families (Figure 1) can be seen as a guide to those substitutions that are favourable. Through this there is reason in claiming that the substitutions that more often occur in a gene family protein are those that result in a sequence similar to other gene family proteins.

The peptide IYYADSVK in Table 20 displays a situation more dire. Here the peptide IYYADSVK is unique for IGHV3-69-1. There are 10 alternative assignments that are possible with a single substitution and 13 from two substitutions. The rate of substitution implicit in the IgFamily program model retains a $\boldsymbol{D_{IGHV3\text{-}69\text{-}1}(LYYADSVK)}$ of 0.911 for the gene family IGHV3-69-1 with the remainder of the density distributed mostly around gene families with one substitution from the IGHV3-69-1 germline.

Table 20 - Although the peptide IYYADSVK is unique for the IGHV3-69-1 germline, there are many alternative assignments with a single substitution.

$\boldsymbol{p}$	$\boldsymbol{q}$	$\boldsymbol{g}$	$\boldsymbol{E_g(p)}$	$\boldsymbol{D_g(p)}$
IYYADSVK	IYYADSVK	IGHV3-69-1	0	0.91054
IYYADSVK	YYADSVK	IGHV3-23	1	0.01020
IYYADSVK	YYADSVK	IGHV3-64	1	0.01020
IYYADSVK	YYADSVK	IGHV3-30	1	0.00921
IYYADSVK	YYADSVK	IGHV3-30-3	1	0.00921
IYYADSVK	YYADSVK	IGHV3-30-5	1	0.00921
IYYADSVK	YYADSVK	IGHV3-33	1	0.00921
IYYADSVK	YYADSVK	IGHV3-43	1	0.00921
IYYADSVK	YYADSVK	IGHV3-66	1	0.00921
IYYADSVK	YYADSVK	IGHV3-53	1	0.00820
IYYADSVK	YADSVK	IGHV3-9	1	0.00441
IYYADSVK	HYADSVK	IGHV3-35	2	0.00012
IYYADSVK	IHHADSVK	IGHV3-29	2	9.34E-05
IYYADSVK	IHHADSVK	IGHV3-30-42	2	9.34E-05
IYYADSVK	YYVDSVK	IGHV3-7	2	8.79E-05
IYYADSVK	YYADSRK	IGHV3-38-3	2	8.79E-05
IYYADSVK	YYVDSVK	IGHV3-52	2	8.79E-05
IYYADSVK	IHHADSVK	IGHV3-30-22	2	8.26E-05
IYYADSVK	YADSVK	IGHV3-20	2	8.24E-05
IYYADSVK	YYADSRK	IGHV3-38	2	7.81E-05
IYYADSVK	IHHADSVK	IGHV3-32	2	7.28E-05
IYYADSVK	YADSVK	IGHV3-11	2	6.52E-05
IYYADSVK	YYADSV	IGHV3-47	2	6.52E-05
IYYADSVK	YADSVK	IGHV3-74	2	5.77E-05
IYYADSVK	HYVDSVK	IGHV3-16	3	1.02E-06

Table 21 and Table 22 show the peptides YYVDSVK and YYADSVK. The peptide YYVDSVK is not a unique peptide and has sequence identity to the IGHV3-7 and IGHV3-52 germlines. The peptide YYADSVK has sequence identity to a total of 12 gene families. Both peptides have a deluge of possible assignments with the consideration of substitutions. Inference on the basis of unique peptides would view these as insufficient evidence for determining the gene family proteins in a sample. Returning to the example sample, consider Table 23. By view of the full body of data it is possible to arrive at an interesting conclusion.

The apparent clonal divergence of the data suggests that the observed peptides are drifting from the germline proteins. In particular the gene families IGHV3-7 and IGHV3-69 have an overlap of peptides possible by substitution. Using the mismatch penalty to investigate the substitution rate, the default rate $\boldsymbol{\alpha_2}$ set as 0.30 results in the assignment of the unique peptides WVANIK and IYYADSVK to their respective gene families as expected. The shared peptide YYVDSVK is assigned to IGHV3-7 with a $\boldsymbol{D_{IGHV3\text{-}7}(YYVDSVK)}$ of 0.475 increasing to a $\boldsymbol{C_{IGHV3\text{-}7}(YYVDSVK)}$ of 0.994 and the shared peptide YYADSVK is assigned to IGHV3-69-1 with a $\boldsymbol{D_{IGHV3\text{-}69\text{-}1}(YYADSVK)}$ of 0.083 increasing to a $\boldsymbol{C_{IGHV3\text{-}69\text{-}1}(YYADSVK)}$ of 0.856. By comparing to Table 21 and Table 22 this is a reasonable assignment. The evidence for IGHV3-7 and IGHV3-69-1 is conferred by the unique peptides WVANIK and IYYADSVK and the assignment of the shared peptides is the most conservative - those with the supporting evidence and least substitutions necessary to explain the data. At the current rate of substitution the amount of evidence to pull an assignment away to a germline sequence with a substitution would be sufficiently large.

What if the rate of substitution was greater than the model currently considers? It seems plausible that the peptide YYADSVK could have come from IGHV3-7, although it would need two substitutions compared to the germline assignment to IGHV3-69-1. With the mismatch penalty $\boldsymbol{\alpha_2}$ set to 0.60 (Table 23) there is little change to the peptides WVANIK and YYVDSVK compared to an $\boldsymbol{\alpha_2}$ of 0.30, however the peptide YYADSVK now has some evidence for a IGHV3-7 origin with a $\boldsymbol{C_{IGHV3\text{-}7}(YYADSVK)}$ of 0.134 while the assignment to IGHV3-69-1 has lowered to a $\boldsymbol{C_{IGHV3\text{-}69\text{-}1}(YYADSVK)}$ of 0.609. Interestingly, even the unique peptide IYYADSVK has a slight change to be from a IGHV3-7 progenitor with a $\boldsymbol{C_{IGHV3\text{-}7}(LYYADSVK)}$ of 0.024. The scenario becomes telling with a mismatch penalty of $\boldsymbol{\alpha_2}$ set to 1.00. The peptides YYADSVK and IYYADSVK are suggested to be from a IGHV3-7 clonal divergence with a $\boldsymbol{C_{IGHV3\text{-}7}(YYADSVK)}$ of 0.569 and a $\boldsymbol{C_{IGHV3\text{-}7}(LYYADSVK)}$ of 0.584. The overall evidence now supports IGHV3-7 and little remains for IGHV3-69-1.

Unique peptides are intuitive to consider as evidence for the gene family proteins present in a sample. However, if the rate of substitution is significant or confounded by the varied nature of immunoglobulin divergence, it is possible that a peptide could be misassigned as a unique germline peptide rather than a divergent peptide from another gene family. Although it is difficult to conclude whether the peptides in this example are from a IGHV3-7 or IGHV3-69 origin - or indeed, from both - it is a key point that relying on unique peptides has the potential for misassignment in general, especially as the inference from the data is carried from a reduced subset of the body of data that is available.

There is more to examine with this idea and it opens insight into the role of immunoglobulins and their diversification. It is suggested here that to assign peptides to gene family germlines and to develop an understanding of the gene family usage in the body, the role of clonal divergence should be studied in unison. Future work should approach the divergence that is observed in the data to inform a model about clonal type peptide groupings and in turn assert which clonal types are representative of a sample.

Table 21 - The peptide YYVDSVK has sequence identity to two gene families.

p	q	g	$E_g(p)$	$D_g(p)$
YYVDSVK	YYVDSVK	IGHV3-7	0	0.47470
YYVDSVK	YYVDSVK	IGHV3-52	0	0.47470
YYVDSVK	HYVDSVK	IGHV3-16	1	0.00568
YYVDSVK	YYADSVK	IGHV3-23	1	0.00453
YYVDSVK	YYADSVK	IGHV3-30	1	0.00453
YYVDSVK	YYADSVK	IGHV3-30-3	1	0.00453
YYVDSVK	YYADSVK	IGHV3-30-5	1	0.00453
YYVDSVK	YYADSVK	IGHV3-33	1	0.00453
YYVDSVK	YYADSVK	IGHV3-43	1	0.00453
YYVDSVK	YYADSVK	IGHV3-53	1	0.00453
YYVDSVK	YYADSVK	IGHV3-66	1	0.00453
YYVDSVK	YYADSVK	IGHV3-69-1	1	0.00453
YYVDSVK	YYADSVK	IGHV3-64	1	0.00402
YYVDSVK	HYADSVK	IGHV3-35	2	5.26E-05
YYVDSVK	YADSVK	IGHV3-20	2	3.87E-05
YYVDSVK	YYADSV	IGHV3-47	2	3.39E-05

Table 22 - The peptide YYADSVK has sequence identity to many gene families.

p	q	g	$E_g(p)$	$D_g(p)$
YYADSVK	YYADSVK	IGHV3-21	0	0.08289
YYADSVK	YYADSVK	IGHV3-23	0	0.08289
YYADSVK	YYADSVK	IGHV3-30	0	0.08289
YYADSVK	YYADSVK	IGHV3-30-3	0	0.08289
YYADSVK	YYADSVK	IGHV3-30-5	0	0.08289
YYADSVK	YYADSVK	IGHV3-33	0	0.08289
YYADSVK	YYADSVK	IGHV3-43	0	0.08289
YYADSVK	YYADSVK	IGHV3-48	0	0.08289
YYADSVK	YYADSVK	IGHV3-53	0	0.08289
YYADSVK	YYADSVK	IGHV3-64	0	0.08289
YYADSVK	YYADSVK	IGHV3-66	0	0.08289
YYADSVK	YYADSVK	IGHV3-69-1	0	0.08289
YYADSVK	HYADSVK	IGHV3-35	1	0.00088
YYADSVK	YYVDSVK	IGHV3-52	1	0.00079
YYADSVK	YYVDSVK	IGHV3-7	1	0.00070
YYADSVK	YADSVK	IGHV3-20	1	0.00066
YYADSVK	YYADSV	IGHV3-47	1	0.00058
YYADSVK	YADSVK	IGHV3-74	1	0.00058
YYADSVK	YADSVK	IGHV3-11	1	0.00052
YYADSVK	YADSVK	IGHV3-9	1	0.00052
YYADSVK	HHADSVK	IGHV3-32	2	8.09E-06
YYADSVK	HYVDSVK	IGHV3-16	2	7.09E-06
YYADSVK	HHADSVK	IGHV3-29	2	7.09E-06

Table 23 - The peptides WVANIK and IYYADSVK are strongly distinct for IGHV3-7 and IGHV3-69-1 respectively. A conservative substitution rate supports association of YYVDSVK to IGHV3-7 and YYADSVK to IGHV3-69-1. However, as the substitution rate is increased, the peptides IYYADSVK and YYADSVK are more readily assigned to IGHV3-7. The overall evidence in the sample supports substitution of the IGHV3-7 germline to mimic peptides originating from IGHV3-69-1.

p	q	g	$\#p$	$E_g(p)$	α_2	$D_g(p)$	$C_g(p)$
WVANIK	WVANIK	IGHV3-7	2	0	0.30	0.999	0.999
WVANIK	WVANIK	IGHV3-69-1	2	5	0.30	0.000	0.000
YYVDSVK	YYVDSVK	IGHV3-7	6	0	0.30	0.475	0.994
YYVDSVK	YYADSVK	IGHV3-69-1	6	1	0.30	0.005	0.006
YYADSVK	YYVDSVK	IGHV3-7	13	1	0.30	0.001	0.012
YYADSVK	YYADSVK	IGHV3-69-1	13	0	0.30	0.083	0.856
IYYADSVK	YYVDSVK	IGHV3-7	4	2	0.30	0.000	0.000
IYYADSVK	IYYADSVK	IGHV3-69-1	4	0	0.30	0.911	0.998

p	q	g	$\#p$	$E_g(p)$	α_2	$D_g(p)$	$C_g(p)$
WVANIK	WVANIK	IGHV3-7	2	0	0.60	0.999	0.999
WVANIK	WVANIK	IGHV3-69-1	2	5	0.60	0.000	0.000
YYVDSVK	YYVDSVK	IGHV3-7	6	0	0.60	0.300	0.940
YYVDSVK	YYADSVK	IGHV3-69-1	6	1	0.60	0.032	0.043
YYADSVK	YYVDSVK	IGHV3-7	13	1	0.60	0.008	0.134
YYADSVK	YYADSVK	IGHV3-69-1	13	0	0.60	0.078	0.609
IYYADSVK	YYVDSVK	IGHV3-7	4	2	0.60	0.004	0.024
IYYADSVK	IYYADSVK	IGHV3-69-1	4	0	0.60	0.308	0.862

p	q	g	$\#p$	$E_g(p)$	α_2	$D_g(p)$	$C_g(p)$
WVANIK	WVANIK	IGHV3-7	2	0	1.00	0.999	0.999
WVANIK	WVANIK	IGHV3-69-1	2	5	1.00	0.000	0.000
YYVDSVK	YYVDSVK	IGHV3-7	6	0	1.00	0.096	0.803
YYVDSVK	YYADSVK	IGHV3-69-1	6	1	1.00	0.062	0.005
YYADSVK	YYVDSVK	IGHV3-7	13	1	1.00	0.032	0.569
YYADSVK	YYADSVK	IGHV3-69-1	13	0	1.00	0.055	0.011
IYYADSVK	YYVDSVK	IGHV3-7	4	2	1.00	0.033	0.584
IYYADSVK	IYYADSVK	IGHV3-69-1	4	0	1.00	0.075	0.014

Future

- De novo and database process comparison

The algorithmic differences between de novo and database peptide assignment are suspected to confer some level of error to spectral assignment. It is not well understood what factors are responsible for incorrect assignment in either case. The IgFamily program stores information relating scan number to peptide assignment. Through this the de novo and database assignments are able to be compared in each method.

- De novo and database process combination

There are instances of de novo assignment where a peptide generates a low confidence assignment but is also supported by the same assignment through the database method. It is possible to compare instances when this occurs to support low confidence de novo assignments. This can be extended by also considering commonly observed misassignments. For example, often the terminal amino acid assignments are incorrect by having the residue locations reversed.

- BLAST internal integration

There are notable limitations with using BLAST through an external command line. Although parameters are able to be directed to BLAST during IgFamily runtime, the modifiable parameters, while thorough, are restrictive. In particular BLAST sets a hard-coded threshold for score output. This would not be an oversight for simple protein identification, but it would be appropriate to build a complete model that assigns some probability to peptide association for each protein. Although the effect of this may not be significant except through certain cases of conjugation it would also serve to satisfy the formal definition of the model.

- BLAST custom substitution matrix

The BLOSUM62 substitution matrix described in the Story is most often used for defining the rate of amino acid substitution. Specifically it describes the relative likelihood of any amino acid change occurring against an expected query amino acid. Although the BLOSUM62 is routinely used for determining sequence similarity, it is may not represent the rate of residue change in immunoglobulins. The divergence from the germline can be used to construct a residue substitution frequency matrix. This could better represent observed clonal divergence when used for analysis. Note that the frequency matrix would necessarily be created by the frequency of observed peptides, and the observation of these peptides is biased by the peptides produced by a particular digest. This would need to considered for greater specificity.

- BLAST custom conservation weighting

The BLOSUM62 substitution matrix described in the Story above assumes that the rate of substitution is spatially uniform - That is, any amino acid is equally likely to be substituted than any other through the substitution matrix. It is known that this is not the case for immunoglobulins, where a relatively greater likelihood of mutation occurs in the hypervariable regions. This could be considered for a more representative rate of substitution.

- Analysis of germline divergence

The clonal divergence of germline immunoglobulins is seen in the overall body of peptide data. It would be a worthwhile study to consider the possible progenitor and divergent peptides. A model would be able to determine the possible gene family germlines if it also had consensus with groupings of peptides based on divergences. Further, the model could both inform and learn from divergence data.

- Measures of abundance

The current measure of peptide abundance is through spectral counting. There are many factors that confound the usefulness of spectral counts as a abundance representation. It could be possible to use measures to determine peptide abundance. An example is fragmentation ion intensity, however chromatography variances and fragmentation efficiency is also a concern. It may be possible to model peptide abundance through a positive control modelling method.

- Further statistical modelling

With an idea of the germline divergence the determination of sample gene family proteins could be bolstered. Divergent peptide groupings would more conifidently assign peptides and help resolve how many gene family proteins there are in a sample. In addition, the Poisson model could have explicit variances that are carried towards the posterior probability. Further, cofactors such as experimental and patient variables could be considered by inclusion of a Dirichlet conjugation to the multinomial.

- Analysis of data generation and reproduction

It can be seen that spectral reproduction from sequential samples are reasonably consistent. Assigned peptides from an earlier sample are often observed in a later sample with predictable levels of corresponding spectra. However, the reproduction is not identical, and variance is seen from run to run. In particular the reproduction of spectra over extended periods of time has not been well studied. Claims on the basis of spectral assignment are reliant on the consistency of independent data generation. It could be valuable to study the production of spectra over subsequent and separated analyses. Peptide assignments could be compared from associated runs, or the generation of the spectra themselves using fragment ion matching. The production of spectra may differ depending on the spatial region of the originating peptide.

- Contaminants report and exclusion list generation

The IgFamily program is able to recognise peptides that originate from contaminant proteins. From these a report is able to be generated with the peptide, mass-to-charge ratio, retention time, charge, theoretical mass, and the associated contaminant protein. In addition an exclusion list is produced in a format suitable for the AB Sciex mass spectrometry instrument software. Reports and exclusion lists are able to be generated from multiple samples by combining the results or by selecting only those peptides that are observed in repeat samples.

- Data filesystem

The IgFamily is able to be executed in both local directory and filesystem mode. However, the filesystem mode currently requires the user to manually accession the necessary files into the filesystem directory. The ISO/IEC TS 18822:2015 filesystem library provides dynamic accessioning of files and would allow rapid scalability for new files and inclusion of earlier data.

- Graphical user interface

The IgFamily program would benefit from a functional graphical user interface. Although the user can adjust parameters through the command line, a user interface would be more accessible particularly to new users. It would also allow the integration of diagrams to aid the user in understanding the program.

- Interactive report generation

The IgFamily program currently produces reports for the user in the HTML markup language. Although fast to develop, the HTML language is limited in interactivity. Interactivity could be developed to enable to user to explore the data. There are two approaches: The HTML report could be supplemented with a collection of JavaScript subroutines. This would allow actions such as spectra to be displayed on selection of peptides or parameters to be adjusted post-hoc. However the JavaScript language does not integrate well with the C++ runtime and a convoluted temporary virtual server would need to created. A more robust solution would be to integrate the analysis results into a graphical user interface that can produce spectra and display graphical information using the OpenGL language.

- Spectral summing integration

It is possible to characterise reproduced spectra by considering fragment ion mass-to-charge ratio, retention time, and fragment ion intensity ratio. Spectra that are determined to be produced from the same peptide are able to be summed together to create a spectra that is on average less prone to fragmentation variation and stochastic noise. The previous IgCompose program was able to sum together reproduced spectra and produced higher quality spectra with less fragment ion mass-to-charge error. In addition, noise peaks are able to be resolved from peptide ion peaks and removed from the spectra. The combination of these processes results in a greatly decreased file size, often as much as a one-fifth reduction. File compression also reduces downstream file analysis time. Although the metric of spectral count is lost, abundance is able to be resolved through another measure, such as fragment ion intensity.

- Dynamic error correction

Following a similar technique to spectral summing, spectra that are with high probability determinable to be produced from the same peptide can establish the mass-to-charge ratio error differential over the course of a mass spectrometry survey. The detected mass-to-charge ratios can then be modified to increase the overall accuracy of a sample. This process would remove the need for ProteinPilot in the conventional workflow.

- PEAKS command line integration

The PEAKS proteomics analysis software is able to be executed through a command line. The necessary de novo or database analysis could be integrated into the IgFamily program workflow. Along with the proposed dynamic error correction, the IgFamily program could automate mass spectrometry analysis from the instrument to the report generation.

Production

Version: IgFamily v0.12.3

Release: 2016-20-11

Language: C++14

Development environment: Microsoft Visual Studio Community 2015

Version control: Git

References

Altschul, S. et al., 1990. Basic local alignment search tool. *Journal of Molecular Biology,* pp. 403-410.

Frost, S. et al., 2015. Assigning and visualizing germline genes in antibody repertoires. *Philosophical Transcactions of the Royal Society B.*

Henikoff, S. & Henikoff, J., 1992. Amino Acid Substitution Matrices from Protein Blocks. *Proceedings of the National Academy of Sciences,* 89(22), pp. 10915-10919.

Karlin, S. & Altschul, S., 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences,* Volume 87, pp. 2264-2268.

Zhang, J. et al., 2012. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular & Cellular Proteomics.*