Clustering Complementarity Determining Regions of T-Cell Receptors

Neerja Thakkar, guided by Professor Chris Bailey-Kellogg
Computer Science Department, Dartmouth College

Abstract

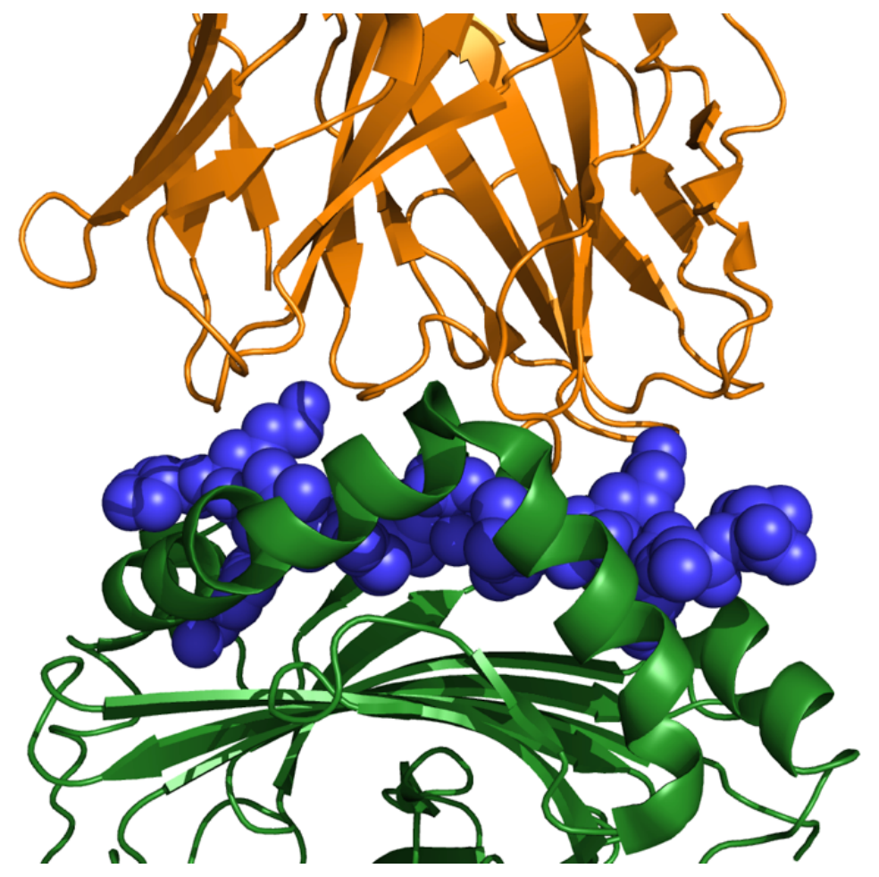
An important component of adaptive immunity in humans lies in the complementarity determining regions (CDRs) of T-cell receptors. CDRs play an important role in distinguishing self vs non-self peptides presented by MHC. CDRs are variable between individuals and the patterns in their sequences have the power to reveal information about genetic diversity in the human immune system. In the Zyvagin paper, differences in twin TCR data were explored. The significance of this exploration lies in the fact that twins have the same genes, so the differences between genetics and exposure/environment can be investigated. This project sought to do an amino acid-level analysis of the same data, in order to characterize differences in the sequence patterns of the CDRs which enable TCRs to recognize different peptides.

Using CDR sequences from TCR repertoires of pairs of twins, a computational method for extracting potentially significant motifs was developed. First, CDR reads are scored pairwise using Smith-Waterman alignment. Then, these scores are used to cluster reads hierarchically. Finally, motifs are extracted using MEME software. For the future, we hope to hone the extraction method and analyze the significance of found motifs.

Background

Complementarity Determining Region of T-Cells

- T-cell: cell of immune system
- Contain T-cell receptor, a molecule which is responsible for recognizing peptide fragments of antigen bound in MHCs (see left figure)
- Much diversity within T-cell receptors, through process of genetic recombination – this diversity is how adaptive immunity works
- Exact genetic impact is unclear, but finding patterns can help us better understand
- Complementarity determining regions (CDRs): part of variable region of TCR
  - Crucial to antigen specificity of TCR



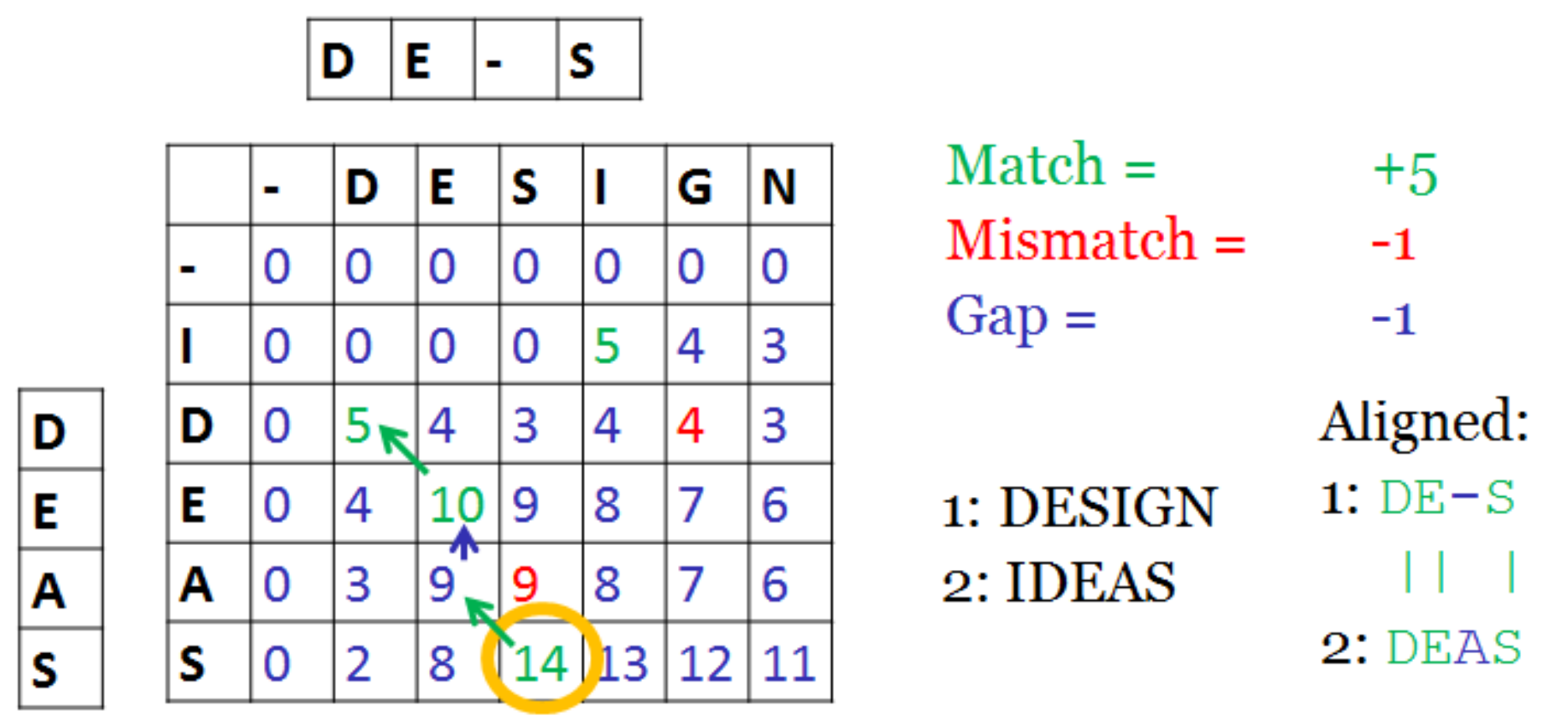
CDRs of TCR (orange) interacting with peptide in the groove of an MHC molecule

Twins Dataset

- 3 pairs of twins, 10000 CDR reads per individual
- Zvyagin paper compares deep TCR repertoires of the twins
  - High-level analysis – found that number of identical CDR3 sequences is increased for twin pairs
- This data was used as a starting point for how to best cluster CDR sequences, looking for amino acid-level specificity differences, as opposed to the higher level analysis used previously
- Questions to answer: by using clustering to analyze CDRs, what information can we extract? How do CDRs compare within an individual, within pairs of twins, and between pairs of twins?

Subsequences

Smith-Waterman Scoring



D	E	-	S

	-	D	E	S	I	G	N
-	0	0	0	0	0	0	0
I	0	0	0	0	5	4	3
D	0	5	4	3	4	4	3
E	0	4	10	9	8	7	6
A	0	3	9	9	8	7	6
S	0	2	8	14	13	12	11

D
E
A
S

Match =	+5
Mismatch =	-1
Gap =	-1

1: DESIGN
2: IDEAS

Aligned:
1: DE-S
|| |
2: DEAS

- Smith-Waterman Algorithm
  Input: a pair of sequences
  Output: a local sequence alignment
- Use an amino acid scoring matrix
- In order to analyze *n* sequences, create an *n* x *n* matrix of scores

Sequences:
1. CAVMDSNYQLIW
2. CAEKSSNTGKLIF
3. CVVNFTGGFKTIF
4. CAASIGLVSNFGNEKLTF
5. CATVGFASGTYKYIF

Score matrix:

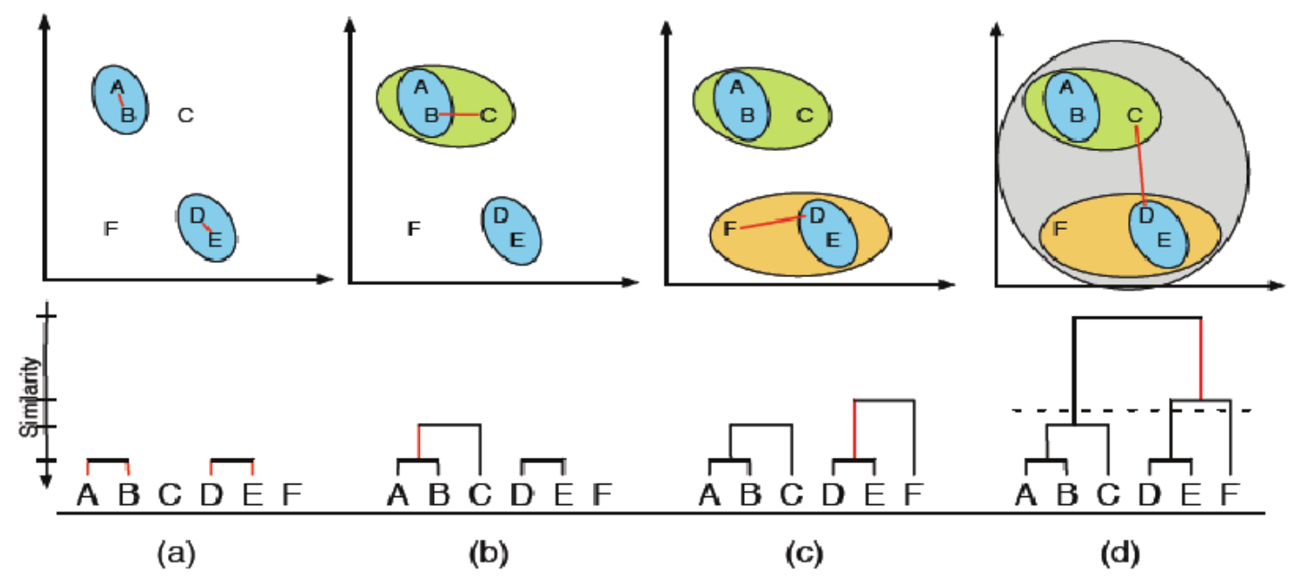
	1	2	3	4	5
1	69	20	13	13	14
2	20	67	15	17	21
3	13	15	72	16	22
4	13	17	16	92	20
5	14	21	22	20	82

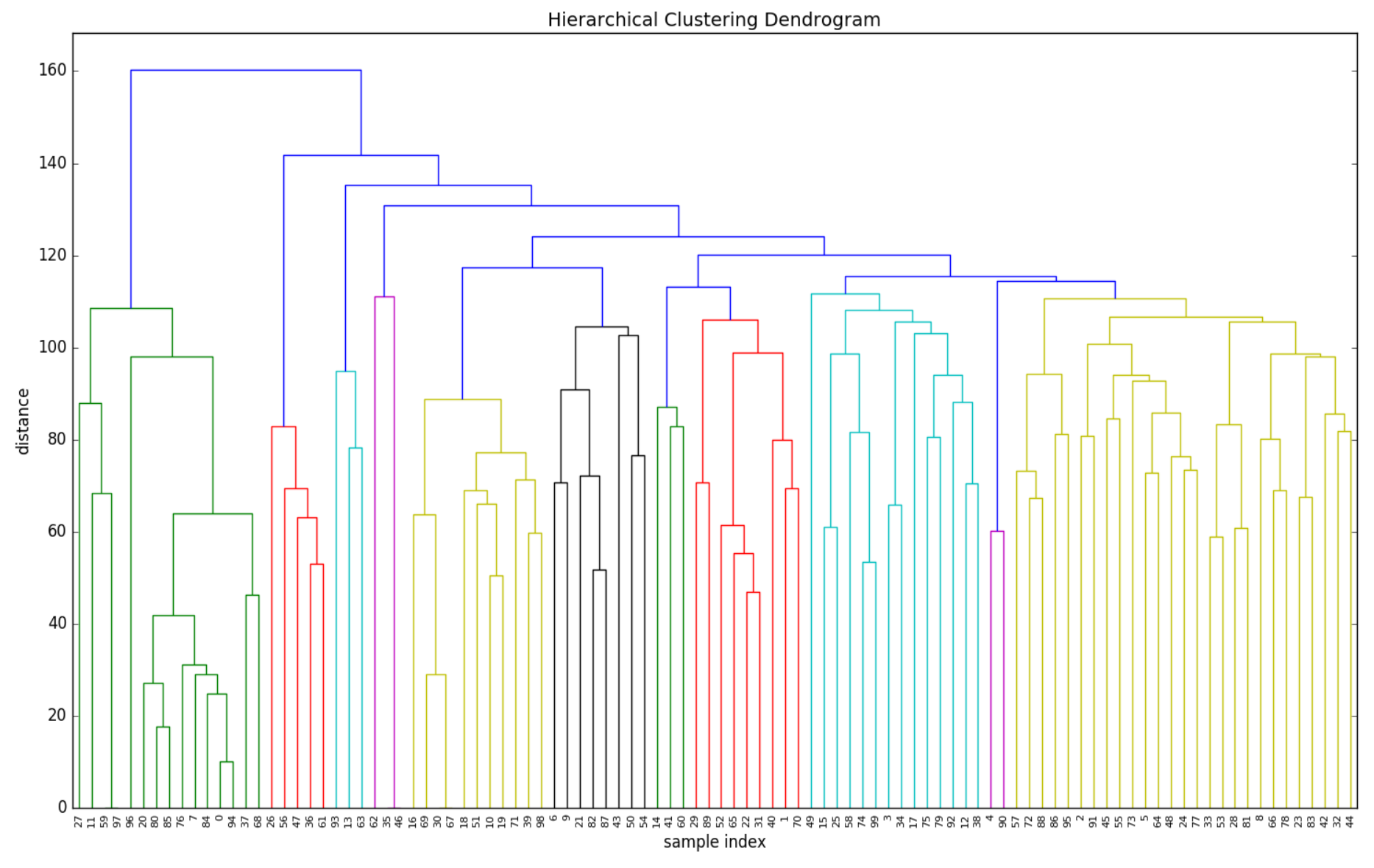
Clustering

Hierarchical Clustering Algorithm

- Clustering: good for unstructured data
- Agglomerative, so "bottom up"
- Build clusters incrementally, forming a dendogram
  - Initialize each sample to its own cluster
  - At each step, merge most similar clusters, using a matrix of scores
- Dendogram gives insight into number of clusters that best fits data

Example: Hierarchical Agglomerative Clustering

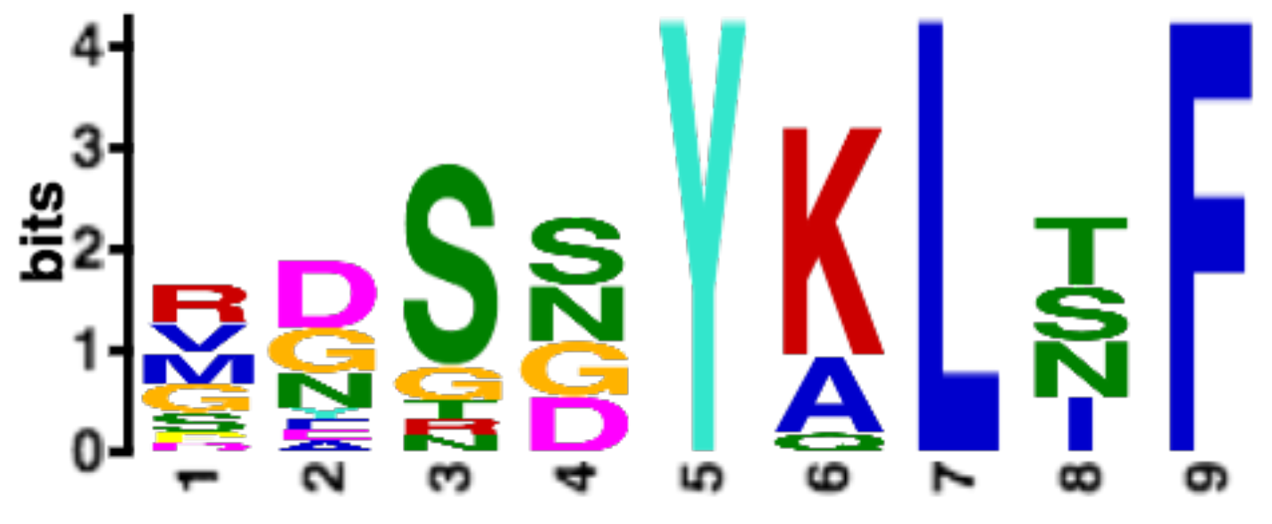




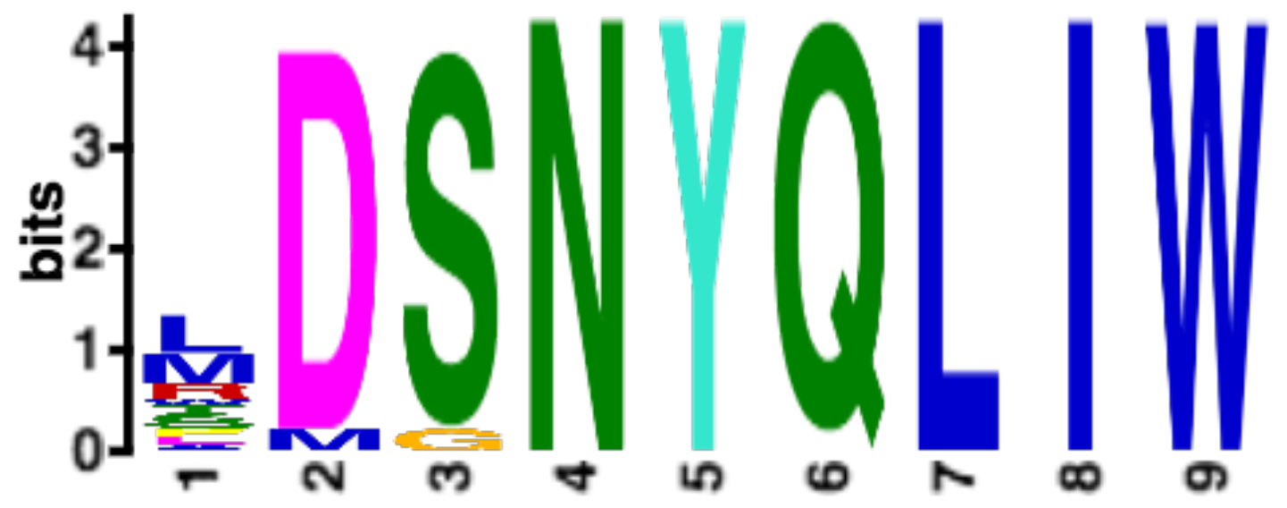
Cluster examples:

Cluster A	Cluster B
AGGSNSGYALNF	AVMDSNYQLIW
AVMDSSYKLIF	AVRDSNYQLIW
AMSEMDSSYKLIF	AATDSNYQLIW
ALSVFNDYKLSF	AVRPDSNYQLIW
AGARSYQLTF	AVSRDSNYQLIW
ATDGTDYKLSF	AVSDSNYQLIW
AVRDGDYKLSF	AALDSNYQLIW
AATPFSGGSNYKLTF	AVEDSNYQLIW
AVRAPYSGYALNF	ASMDSNYQLIW
VVRNSGYALNF	AVLDSNYQLIW
AVRDGDYKLSF	VVSAMGNYQLIW
VVRMDSSYKLIF	AGMDSNYQLIW
VVRNSGYALNF	AVLDSNYQLIW
ALSEAGGSGGSNYKLTF	AVLDSNYQLIW
AVIVGSNYKLTF	AVRDVWDSNYQLIW
AVNQAVESNYKLTF	AVLDSNYQLIW
AESDSSYKLIF	AVMDSNYQLIW
	AVQDSNYQLIW

Motif Extraction



Extracted from Cluster A



Extracted from Cluster B

- Use MEME motif-based sequence analysis
- Motif: a recurring, fixed-length pattern found in sequences
- Probabilistic model for motif discovery
- Generates matrix of probabilities, which is then used to score other sequences based on how well it matches the motif
- Score reads to examine motif occurrences within an individual and between/within twin pairs

Conclusions and Further Directions

- Learned about data and what methods for analysis are potentially useful
- Inconclusive results, but hope to apply analysis of motifs to more data in the future

Two goals for the future:

1. Refine clustering
   - Explore other possibilities besides MEME
   - Remove C-terminal amino acids from analysis
   - Hone in best number of clusters
2. Apply analysis to more data
   - Compare between twins
   - Compare between pairs of twins
   - Get access to larger data sets

Acknowledgements

- Sophomore Science Research Scholars – funding
- Key paper, data: Zvyagin IV, Pogorelyy MV, Ivanova ME, Komech EA, Shugay M, Bolotin DA, Shelenkov AA, Kurnosov AA, Staroverov DB, Chudakov DM, Lebedev YB, Mamedov IZ. Distinctive properties of identical twins' TCR repertoires revealed by high-throughput sequencing. PNAS. 2014;111:5980–5985. doi: 10.1073/pnas.1319389111.