

01 Nielsen et al. (**16**) proposed an advanced motif sampler method based on the
02 Gibbs sampling technique, which efficiently samples the possible alignment
03 space of binder sequences. For each alignment, a log-odds weight matrix was
04 calculated for the identified binding core subsequences. This matrix serves
05 as the position-specific scoring matrix for the computation of a score for a
06 nonamer.

07 Motivated by a text mining model designed for building a classifier from
08 labeled and unlabeled examples, Murugan and Dai (**17**) developed an iterative
09 supervised learning model for the prediction of MHC class II-binding peptides.
10 The iterative learning model, based on LP, enables the use of nonbinder infor-
11 mation for the detection of the binding cores from a set of putative binding
12 cores and for the construction of the predictor simultaneously. The outcome of
13 this predictor is a position-specific weight matrix that can score amino acids at
14 each position of a nonamer.

15 16 **2. Materials**

- 17 1. A data set that includes binding and nonbinding peptides for a specific MHC class II
18 allele. The recommended size of binders is above 100. Any in-house peptide set can
19 be used. If the number of peptides is not sufficient, peptides from databases such
20 as AntiJen (**18**) and MHCBN (**19**) can be added for training. For some alleles, the
21 number of nonbinders may be extremely small. In this case, the random sequences
22 can be added (*see Note 1*).
- 23 2. Predictors that can score the binding ability for each individual peptide (*see Note 2*).

24 25 26 **3. Methods**

27 The Bayesian predictor is trained based on the prediction outcome obtained
28 from each individual predictor for a set of training peptides. The system is
29 flexible to incorporate results from any number of predictors. Suppose that the
30 number of predictors is m . In general, the requirement for each predictor is
31 the generation of a score for a given peptide sequence. This score of a peptide
32 is designated as the highest value among all scores that are assigned to the
33 overlapping 9 mer of the peptide by a predictor. A peptide is predicted as
34 a binder (resp. nonbinder) if this score is above (resp. below) a prescribed
35 threshold value. The steps for building a Bayesian predictor are as follows:

- 36 1. Prepare a training data set. Any peptide sequence with length less than nine residues
37 or with undetermined residues in certain positions should be discarded.
- 38 2. Reduce the redundancy in the data set. This step is to prevent overestimation of
39 the performance of a predictor. After the reduction, there should be no two peptide

- 01 sequences in the set with sequence identity >90% over an alignment of length at
 02 least nine residues.
- 03 3. Obtain a predictive score for each peptide in the training set (including binding and
 04 nonbinding sequences) from each individual predictor. These scores form the input
 05 set from which a Bayesian predictor can be built.
- 06 4. Determine a set of threshold values that produce distinct pairs of sensitivity and
 07 specificity (*see Note 3*). This procedure should be performed for each predictor
 08 on the training set. Upon the completion of this step, a set of threshold values for
 09 predictor j is obtained, say $\delta^j = (\delta_1^j, \dots, \delta_{t_j}^j)$, $j = 1, \dots, m$, where t_j is the number
 10 of possible threshold values with the above property for predictor j .
- 11 5. Determine the best combination $\delta^* = (\delta^{*1}, \dots, \delta^{*m})$ of threshold values, where each
 12 δ^{*j} ($j = 1, \dots, m$) is the selected threshold value for predictor j . This combination
 13 can be determined by finding the highest average area under receiver operating
 14 characteristic curve (AROC) (*see Note 3*) value for the Bayesian predictor with a
 15 k -fold cross-validation procedure described as follows:
- 16 a. For each combination of threshold values $\delta_{i_1}^1, \dots, \delta_{i_m}^m$ set up a prediction
 17 outcome table for the $(k-1)$ -folds of the training peptides (*see Note 4*), where
 18 $\delta_{i_j}^j$ is the i_j th threshold value for predictor j , $j = 1, \dots, m$ and $i_j = 1, \dots, t_j$.
 19 This table is of size $n \times m$, where n is the number of peptides in the training
 20 folds. The outcome obtained from predictor j for a peptide is denoted by a
 21 binary number f_j : $f_j = 1$ if the peptide is predicted as binder, $f_j = 0$ otherwise.
 22 Accordingly, the prediction outcome obtained from the m predictors for each
 23 peptide will be coded by a binary string $f_1 f_2 \dots f_m$.
- 24 b. Build the probability table for the Bayesian predictor from the $n \times m$ table
 25 described above. Let y_i denote the label of each peptide: $y_i = 1$ if it is a binder
 26 and $y_i = -1$ if it is not a binder. The probabilities for each value f_j of the m
 27 features for the binder class and the nonbinder class are computed as follows:

AQ3

$$p(f_j = 1 | \text{binder class}) = \frac{\sum_{i: y_i = 1} I(f_{ij} = 1)}{\text{total number of binders}}, \quad j = 1, \dots, m,$$

$$p(f_j = 0 | \text{binder class}) = \frac{\sum_{i: y_i = 1} I(f_{ij} = 0)}{\text{total number of binders}}, \quad j = 1, \dots, m,$$

$$p(f_j = 1 | \text{nonbinder class}) = \frac{\sum_{i: y_i = -1} I(f_{ij} = 1)}{\text{total number of nonbinders}}, \quad j = 1, \dots, m, \text{ and}$$

$$p(f_j = 0 | \text{nonbinder class}) = \frac{\sum_{i: y_i = -1} I(f_{ij} = 0)}{\text{total number of nonbinders}}, \quad j = 1, \dots, m,$$

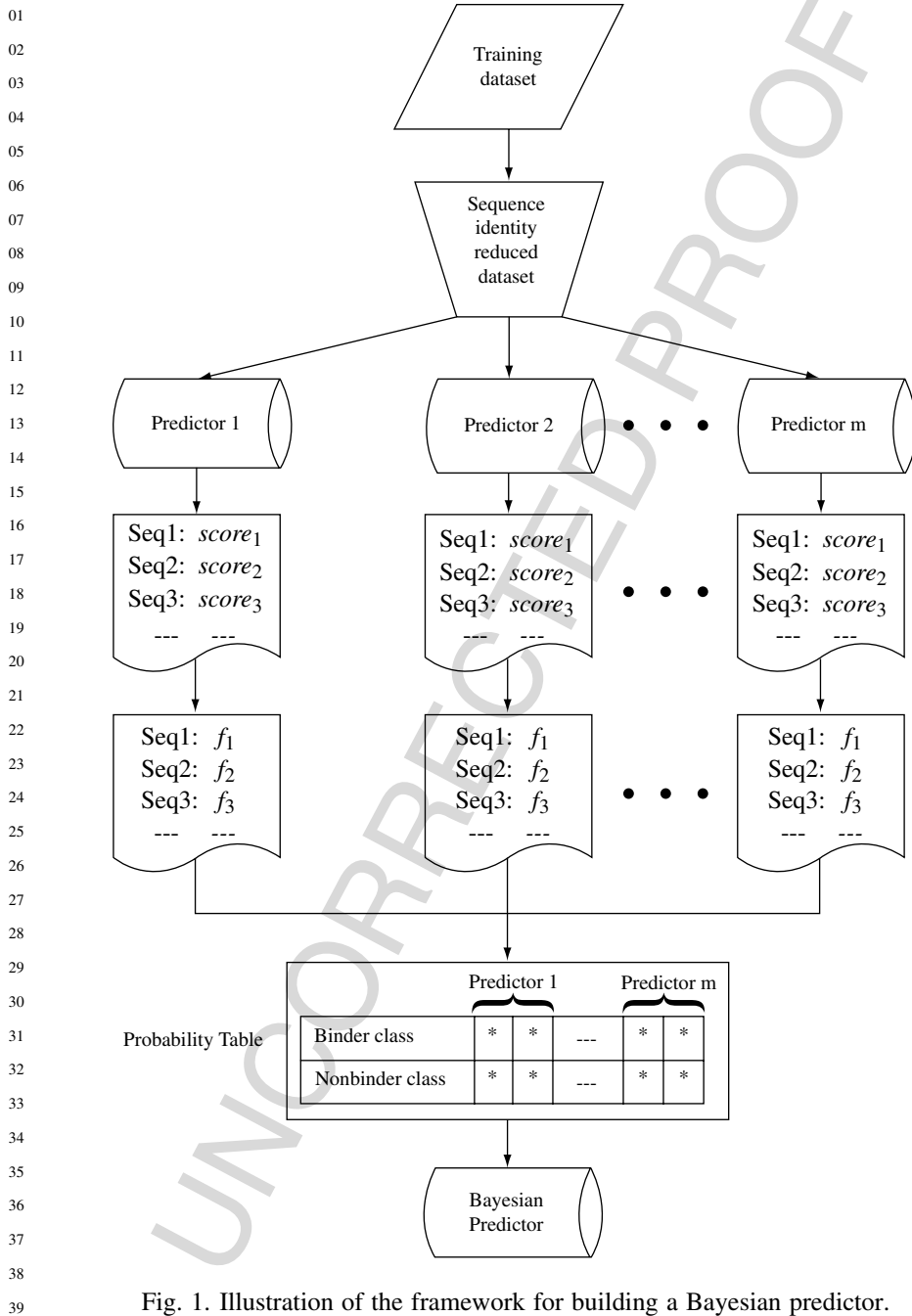


Fig. 1. Illustration of the framework for building a Bayesian predictor.

References

- 01 1. Flower, D. R. (2004) Vaccines in silico – the growth and power of immunoinfor-
02 matics. *The Biochemist* **26**, 17–20.
- 03 2. De Groot, A. S. and Berzofsky, J. A. (2004) From genome to vaccine – new
04 immunoinformatics tools for vaccine design. *Methods* **34**, 425–428.
- 05 3. Parham, P. (2005) *The Immune System*. Garland Science, New York, NY.
- 06 4. Castellino, F., Zhong, G., and Germain, R. N. (1997) Antigen presentation by
07 MHC class II molecules: invariant chain function, protein trafficking, and the
08 molecular basis of diverse determinant capture. *Hum. Immunol.* **54**, 159–169.
- 09 5. Sette, A., Buus, S., Appella, E., Smith, J. A., Chesnut, R., Miles, C., Colon, S. M.,
10 and Grey, H. M. (1989) Prediction of major histocompatibility complex binding
11 regions of protein antigens by sequence pattern analysis. *Proc. Natl. Acad. Sci.*
12 *U.S.A.* **86**, 3296–3300.
- 13 6. Max, H., Halder, T., Kropshofer, H., Kalbus, M., Muller, C. A., and Kalbacher,
14 H. (1993) Characterization of peptides bound to extracellular and intracellular
15 HLA-DR1 molecules. *Hum. Immunol.* **38**, 193–200.
- 16 7. Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A., and Stevanovic, S.
17 (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immuno-*
18 *genetics* **50**, 213–219.
- 19 8. Borrás-Cuesta, F., Golvano, J., García-Granero, M., Sarobe, P., Riezu-Boj, J.,
20 Huarte, E., and Lasarte, J. (2000) Specific and general HLA-DR binding motifs:
21 comparison of algorithms. *Hum. Immunol.* **61**, 266–278.
- 22 9. Sturniolo, T., Bono, E., Ding, J., Radrizzani, L., Tuereci, O., Sahin, U.,
23 Braxenthaler, M., Gallazzi, F., Protti, M. P., Sinigaglia, F., and Hammer, J. (1999)
24 Generation of tissue-specific and promiscuous HLA ligand databases using DNA
25 microarrays and virtual HLA class II matrices. *Nat. Biotechnol.* **17**, 555–561.
- 26 10. Kato, R., Noguchi, H., Honda, H., and Kobayashi, T. (2003) Hidden Markov
27 model-based approach as the first screening of binding peptides that interact with
28 MHC class II molecules. *Enzyme Microb. Technol.* **33**, 472–481.
- 29 11. Brusci, V., Rudy, G., Honeyman, G., Hammer, J., and Harrison, L. (1998)
30 Prediction of MHC class II-binding peptides using an evolutionary algorithm and
31 artificial neural network. *Bioinformatics* **14**, 121–130.
- 32 12. Nielsen, M., Lundegaard, C., Worning, P., Lauemoller, S. L., Lamberth, K.,
33 Buus, S., Brunak, S., and Lund, O. (2003) Reliable prediction of T-cell epitopes
34 using neural networks with novel sequence representations. *Protein Sci.* **12**,
35 1007–1017.
- 36 13. Bhasin, M. and Raghava, G. P. (2004) SVM based method for predicting
37 HLA-DRB1*0401 binding peptides in an antigen sequence. *Bioinformatics* **20**,
38 421–423.
- 39 14. Theodoridis, S. and Koutroumbas, K. (1999) *Pattern Recognition*. Academic Press,
San Diego, CA.

- 01 15. Singh, H. and Raghava, G. P. (2001) ProPred: prediction of HLA-DR binding
02 sites. *Bioinformatics* **17**, 1236–1237.
- 03 16. Nielsen, M., Lundegaard, C., Worning, P., Hvid, C. S., Lamberth, K., Buus, S.,
04 Brunak, S., and Lund, O. (2004) Improved prediction of MHC class I and class II
05 epitopes using a novel Gibbs sampling approach. *Bioinformatics* **20**, 1388–1397.
- 06 17. Murugan, N. and Dai, Y. (2005) Prediction of MHC class II binding peptides
07 based on an iterative learning model. *Immunome Res.* **1**, 6.
- 08 18. Toseland, C. P., Clayton, D. J., McSparron, H., Hemsley, S. L., Blythe, M. J.,
09 Paine, K., Doytchinova, I. A., Guan, P., Hattotuagama, C. K., and Flower, D. R.
10 (2005) AntiJen: a quantitative immunology database integrating functional,
11 thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res.* **1**, 4.
- 12 19. Bhasin, M., Singh, H., and Raghava, G. P. (2003) MHCBN: a comprehensive
13 database of MHC binding and non-binding peptides. *Bioinformatics* **19**, 665–666.
- 14 20. Swets, J. A. (1988) Measuring the accuracy of diagnostic systems. *Science* **240**,
15 1285–1293.
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39