The Immune Epitope Database and Analysis Resource

Sette A[1], Bui HH[1], Sidney J[1], Bourne P[2], Buus S[3], Fleri W[1], Kubo R[1,4], Lund O[5], Nemazee D[6], Ponomarenko JV[2], Sathiamurthy M[1], Stewart S[1], Way S[1], Wilson SS[1], and Peters B[1]

[1] La Jolla Institute of Allergy and Immunology, 9420 Athena Circle,
La Jolla, CA 92037, USA
[2] San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive,
La Jolla, CA 92093, USA
[3] University of Copenhagen, Copenhagen, DK-2200, Denmark
[4] Gemini Science, 9420 Athena Circle, La Jolla, CA 92037, USA
[5] Center for Biological Sequence Analysis, BioCentrum-DTU, Building 208, Technical
University of Denmark, DK-2800 Kgs. Lyngby, Denmark
[6] The Scripps Research Institute, 10555 North Torrey Pines Road, La Jolla, CA 92037, USA
[7] Science Applications International Corporation, San Diego, CA 92121

Abstract. Epitopes are defined as the molecular structures interacting with specific receptors of the immune system such as antibodies, MHC, and T cell receptor molecules. The Immune Epitope Database and Analysis Resource (IEDB, http://www.immuneepitope.org) is a database specifically devoted to immune epitope data. The database is populated with intrinsic and context-dependent epitope data curated from the scientific literature by immunologists, biochemists, and microbiologists. An analysis resource is linked to the database which hosts various bioinformatics tools to analyze epitope data as well as to predict de novo epitopes. The availability of the IEDB will facilitate the exploration of immunity to infectious diseases, allergies, autoimmune diseases, and cancer. The utility of the IEDB was recently demonstrated through a comprehensive analysis of all current information regarding antibody and T cell epitopes derived from influenza A and determining possible cross-reactivity among H5N1 avian flu and human flu viruses.

1 Introduction

Epitopes are defined as the molecular structures interacting with specific receptors of the immune system such as antibodies, MHC, and T cell receptor molecules. Knowledge of the epitopes involved in the immune response is critical to detect, monitor, and design therapies to fight infectious diseases as well as allergies, autoimmunity and cancer. A vast amount of epitope-related information is available, ranging from epitope binding affinities for their receptors, to cellular and humoral responses, to data analyzing correlates of protection or immune pathology. We have developed a central resource that captures this information, allowing users to connect realms of knowledge currently separated and difficult to access. This new initiative, "The Immune Epitope Database and Analysis Resource", became available to the public in a beta version on 15 February 2006 (http://www.immuneepitope.org) [1, 2].

J.C. Rajapakse, L. Wong, and R. Acharya (Eds.): PRIB 2006, LNBI 4146, pp. 126 – 132, 2006.

The priorities for inclusion in the database are epitopes from category A-C pathogens such as influenza, SARS and poxviruses and emerging/re-emerging infectious diseases. However, we anticipate that all immune epitope data will eventually be curated. B and T cell epitopes recognized in humans, non-human primates and laboratory animals are all considered within the scope of the project. Accordingly, we estimate that about 100,000 different literature records to be relevant. In addition, we expect to host a large volume of direct data submissions from various NIH-sponsored research contracts. Curation and query strategies have been developed to enable effective handling of this large amount of highly context dependent information.

An analysis resource is linked to the database that hosts various bioinformatic tools to analyze epitope data (including, for example, population coverage and sequence conservation), as well as tools to predict epitope cellular processing, binding to MHC, and recognition by T cell receptors and antibody molecules. In this context we observed that a large number of predictive tools related to epitopes exist in the literature, and new ones are continuously being developed. Evaluating the performance of existing and newly developed tools will be an on-going effort for the IEDB team.

2 Database Structure and Design

The IEDB has been developed as a web-accessible database using an industry standard software design. The IEDB application is a Model View Controller (MVC) (http://java.sun.com/blueprints/guidelines/designing_enterprise_applications_2e/web-tier/web-tier5.html) style Enterprise Java (J2EE) application with a relational database management system (Oracle 10g) data repository. The system architecture is divided into two websites, three application tiers, and two physical servers (See Figure 1). The application was constructed using existing Java frameworks and commercial products to create the infrastructure allowing the development team to concentrate on the novel functionality that was required. The enterprise architecture is flexible, extensible, scalable, and proven.

Each epitope is associated with intrinsic and extrinsic features. Intrinsic features are those associated with its sequence and structure, while extrinsic features are context-dependent attributes dependent upon the specific experimental or natural environment context. Contextual information includes the species, health, and genetic makeup of the host, the immunization route and dose, and the presence of adjuvants. In order to describe an immune response associated with a specific epitope, both the intrinsic and extrinsic (context-dependent) features need to be taken into account. This immunological perspective has been a guiding principle in organizing the data in the IEDB [1-3].

The hierarchal nature of the data has been captured in an epitope-centric, ontology-like data structure [4], developed in a top-down process where each class in a domain and its properties were defined before building the hierarchy. The primary classes of the IEDB consist of *Reference*, *Epitope Structure*, *Epitope Source*, *MHC Binding*, *MHC Ligand Elution*, *T Cell Response*, and *B Cell Response* (See Figure 2). The *Epitope* class encapsulates all the individual concepts identified. In turn, the

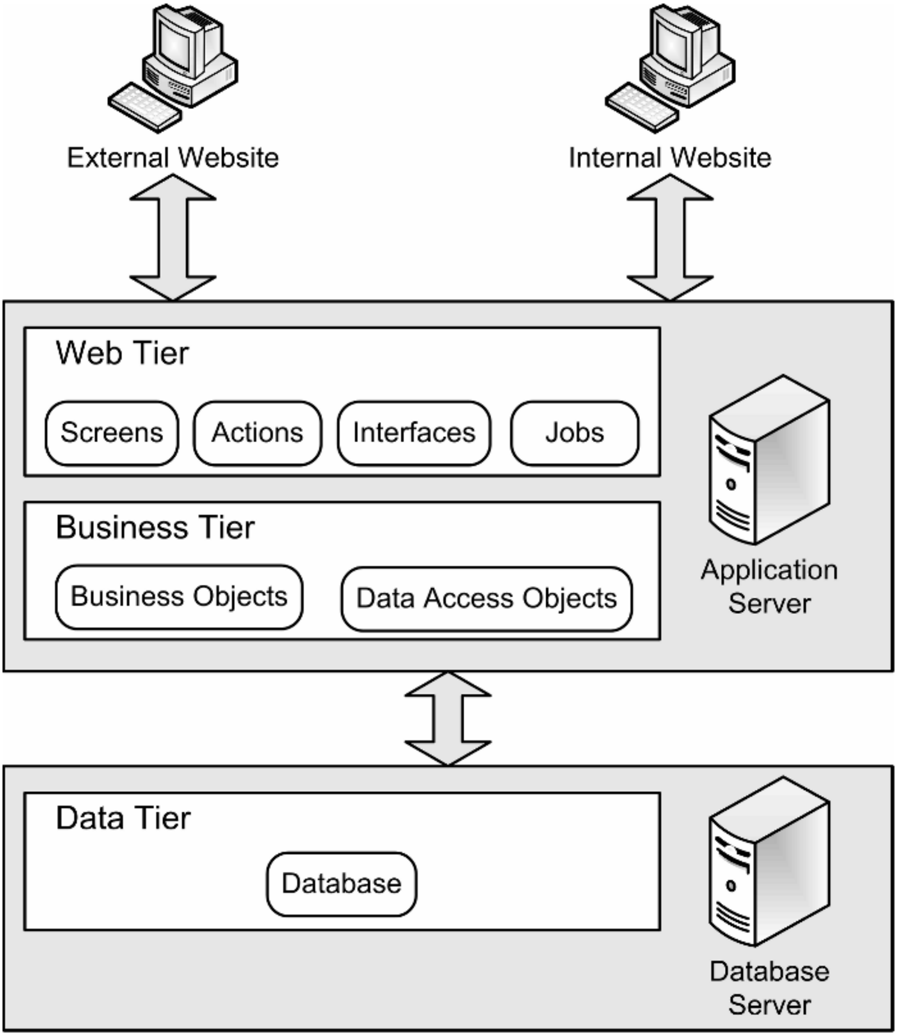

Fig. 1. Logical to Physical Tier Map

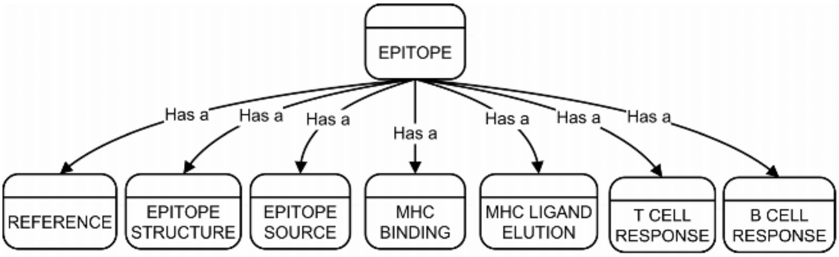

Fig. 2. Detailed classification of Epitope class showing its properties

individual concepts are related to other classes. The primary relationships have a subclass relationship or use a property (denoted in the figure by the arcs labeled "Has a") that has a value restriction.

3 Populating the Database

To identify and extract relevant data from the scientific literature in an efficient and accurate manner, novel formalized curation strategies were developed, enabling the processing of a large volume of context-dependent data. This process is multi-step, involving an automated PubMed query, a manual abstract scan to select potentially relevant references, followed by methodical analysis of the selected references, and finally the manual curation of papers deemed relevant to the scope of the database by a team of dedicated curators with expertise in the areas of biochemistry, microbiology and immunology. Once the manual curation of a reference is complete, the curation is reviewed by an independent group of immunologists and structural biologists, thus, integrating experts and data curators to optimize quality, consistency and uniformity.

To facilitate accurate translation of the information contained in the literature into the structured format of the database, we developed a Curation Manual and Data Ontology. These documents are designed to provide a consistent set of rules, definitions, and guidelines regarding the strategies and procedures for capturing, annotating and introducing data from the literature into the IEDB. Additionally, feedback from external experts in the fields of immunology and infectious diseases has been sought on an ongoing basis in order to improve both the database structure and curation practices. In this way, complex experimental data are captured in a consistent and accurate manner.

Management of the curation of a large number of references by a team of curators and reviewers required the development of a formal tracking system. All transactions and comments pertaining to each reference are tracked to provide details of the progress of each curated paper from selection of the manuscript to final incorporation of the data into the IEDB. As of May 2006, over 1900 references have been manually curated.

4 Analysis Resource

An analysis resource is linked to the database allowing users to analyze epitope data as well as to predict de novo epitopes. For example, one tool predicts the *population coverage* for a user-prescribed set of T-cell epitopes [5], i.e. the fraction of an ethnic population likely to respond to one or more epitopes in the set. This is done by relating the known MHC restrictions of the epitopes to frequencies of the MHC alleles in different populations, and calculating the total population covered assuming linkage equilibrium between MHC loci. An illustration of the utility of this tool is that it allows a user to detect if a set of epitopes, which may be intended for use as a diagnostic tool or vaccine, has ethnically skewed or balanced population coverage.

Another tool calculates the protein sequence *conservation* of epitopes. For a given starting sequence and threshold of sequence identity, the tool calculates the fraction of proteins containing the epitope. Focusing on epitopes that are conserved at a high level of sequence identity is specifically important for RNA viruses, which show a large degree of sequence variability between different isolates or strains.

To predict the presence of *antibody epitopes* in protein sequences, a number of previously existing amino acid scale-based tools have been implemented from the literature. Although these particular tools have been shown to underperform [6], they represent the current state-of-the-art in antibody epitope prediction and do provide a benchmark for future tool development. Our intent is to devote significant efforts towards the development of improved tools for the prediction of antibody epitopes.

The most extensively tested predictions at present are those describing *peptide binding to MHC class I molecules*. The ability of a peptide to bind an MHC molecule is a necessary requirement for it to be recognized by T-cells. As MHC molecules are very specific in their binding preference, these predictions provide a powerful means to scan entire pathogens for T-cell epitope candidates. Three separate prediction methods were implemented, two of them based on scoring matrices (ARB [7] and SMM [8]) and one based on an artificial neural network [9, 10]. The three methods were compared using five-fold cross validation on a large dataset comprising nearly 50,000 data points, in which the neural network based predictions outperformed the other two [11]. The complete benchmark dataset used in this evaluation is available at http://mhcbindingpredictions.immuneepitope.org/, and we encourage developers to use these data in training and testing their own tools.

In addition to the tools described above, tools for predicting MHC class II epitopes [7], proteasomal cleavage [12] and TAP transport [13] have also been implemented and will be evaluated in a similar manner to the MHC class I binding predictions. Also, for epitopes with a 3D structure available in the PDB, an *epitope viewer* has been developed displaying the epitope structure and its immune receptor interactions.

5 Curating and Analyzing Influenza a Epitope Data

As pointed out in a recent Nature editorial, the fight against flu is undermined by "the lack of an accessible store of information" [14]. Besides outbreaks and sequencing data, information is also lacking regarding influenza epitopes. This knowledge is crucial to predict potential cross-reactive immunity and coverage of new strains by vaccines and diagnostic candidates. To demonstrate the features of IEDB and in response to the global spread of highly virulent H5N1 influenza viruses, we have performed an analysis of influenza A epitope information to: 1) compile all current information regarding antibody and T cell epitopes derived from influenza A and 2) determine possible cross-reactivity among H5N1 avian flu and human flu virus.

To compile all information available in the literature relating to influenza epitopes, we inspected over 2000 references, and more than 400 were added to the IEDB after detailed curation. An assessment of these curated records revealed that approximately 600 different epitopes, derived from 58 strains, recognized in 8 different hosts and derived from all flu proteins have been identified and reported in the literature, including several conserved epitopes and a small number of protective ones. The latter are of particular interest as they may confer cross-reactive protection against influenza strains of the avian H5N1 subtype. Significantly, however, this analysis made apparent the fact that: 1) few protective antibody and T cell epitopes are reported in the literature; 2) there is a paucity of antibody epitopes in comparison to T cell epitopes; 3) the number of animal hosts from which the epitopes were defined is limited; 4) the number of epitopes reported for avian influenza strains/subtypes is limited, 5) the number of epitopes reported from proteins other than hemagglutinin (HA) and nucleoprotein (NP) is limited. In summary, this analysis provides a unique resource to evaluate existing data and to aid efforts in guarding against seasonal and pandemic flu outbreaks.

6 Conclusion

The IEDB is an initiative focused on creating large volumes of complex context-dependent immunological data paired with relevant analytical tools. The project should facilitate basic research, as well as the development of new vaccines and diagnostics. The experience gained in the process of developing and operating the IEDB will be of value in the development and integration of other biological databases capturing clinical, immunological, genomic and cellular biology knowledge.

Acknowledgments. This work was supported by the National Institutes of Health Contract HHSN26620040006C.

References

1. Peters B, Sidney J, Bourne P, Bui HH, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O et al. The design and implementation of the immune epitope database and analysis resource. Immunogenetics 2005.
2. Peters B, Sidney J, Bourne P, Bui HH, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O et al. The immune epitope database and analysis resource: from vision to blueprint. PLoS Biol 2005, 3(3):e91.
3. Sette A, Fleri W, Peters B, Sathiamurthy M, Bui HH, Wilson S: A roadmap for the immunomics of category A-C pathogens. Immunity 2005, 22(2):155-161.
4. Sathiamurthy M, Peters B, Bui HH, Sidney J, Mokili J, Wilson SS, Fleri W, McGuinness DL, Bourne PE, Sette A. An ontology for immune epitopes: application to the design of a broad scope database of immune reactivities. Immunome Res. 2005 Sep 20;1(1):2.
5. Bui HH, Sidney J, Dinh K, Southwood S, Newman MJ, et al. (2006) Predicting population coverage of T-cell epitope-based diagnostics and vaccines. BMC Bioinformatics 7: 153.
6. Blythe MJ, Flower DR (2005) Benchmarking B cell epitope prediction: underperformance of existing methods. Protein Sci 14: 246-248.
7. Bui HH, Sidney J, Peters B, Sathiamurthy M, Sinichi A, et al. (2005) Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. Immunogenetics 57: 304-314.
8. Peters B, Sette A (2005) Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. BMC Bioinformatics 6: 132.
9. Buus S, Lauemoller SL, Worning P, Kesmir C, Frimurer T, et al. (2003) Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach. Tissue Antigens 62: 378-384.

10. Nielsen M, Lundegaard C, Worning P, Lauemoller SL, Lamberth K, et al. (2003) Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci 12: 1007-1017.
11. Peters B, Bui HH, Frankild S, Nielsen M, Lundegaard C, Kostem E, Basch D, Lamberth K, Harndahl M, Fleri W, Wilson S, Sidney J, Lund O, Buus S and Sette A. (2006) A Community Resource Benchmarking Predictions of Peptide Binding to MHC-I Molecules. PLoS Comp Bio, in press.
12. Tenzer S, Peters B, Bulik S, Schoor O, Lemmel C, et al. (2005) Modeling the MHC class I pathway by combining predictions of proteasomal cleavage,TAP transport and MHC class I binding. Cell Mol Life Sci 62: 1025-1037.
13. Peters B, Bulik S, Tampe R, Van Endert PM, Holzhutter HG (2003) Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. J Immunol 171: 1741-1749.
14. Dreams of flu data. Nature 440, 255-6 (2006)