THE DIFFERENCE OF BREAKING THROUGH TO NEW FRONTIERS IN SINGLE-CELL ANALYSIS

SPEARHEADING NEW DISCOVERIES WITH THE FIRST-OF-ITS-KIND BD FACSYMPHONY™ S6 CELL SORTER. The BD FACSymphony™ S6 cell sorter offers analysis of up to 60 parameters with 6-way sorting—powering deeper and richer scientific insights than ever before. Now, researchers can characterize and sort single-cell populations to unearth new insights on the cutting edge of science. With scalable and upgradeable architecture, the BD FACSymphony™ S6 cell sorter can grow with you as your scientific needs change. Discover a cell sorter that helps you push the boundaries of scientific inquiry. **Discover the new BD.**

Learn how you can advance your research >

BD

Immunology
The Journal of cells, molecules, systems and technologies

British Society for immunology

IMMUNOLOGY SPOTLIGHT

The immune epitope database: a historical retrospective of the first decade

Nima Salimi,* Ward Fleri, Bjoern Peters and Alessandro Sette
La Jolla Institute for Allergy & Immunology, La Jolla, CA, USA

doi:10.1111/j.1365-2567.2012.03611.x
Received 21 May 2012; revised 23 May 2012; accepted 28 May 2012.

*Correspondence: Nima Salimi, La Jolla Institute for Allergy & Immunology, 9420 Athena Circle, La Jolla, CA 92037, USA.
Email: nsalimi@liai.org
Senior author: Nima Salimi

Summary

As the amount of biomedical information available in the literature continues to increase, databases that aggregate this information continue to grow in importance and scope. The population of databases can occur either through fully automated text mining approaches or through manual curation by human subject experts. We here report our experiences in populating the National Institute of Allergy and Infectious Diseases sponsored Immune Epitope Database and Analysis Resource (IEDB, http://iedb.org), which was created in 2003, and as of 2012 captures the epitope information from approximately 99% of all papers published to date that describe immune epitopes (with the exception of cancer and HIV data). This was achieved using a hybrid model based on automated document categorization and extensive human expert involvement. This task required automated scanning of over 22 million PubMed abstracts followed by classification and curation of over 13 000 references, including over 7000 infectious disease-related manuscripts, over 1000 allergy-related manuscripts, roughly 4000 related to autoimmunity, and 1000 transplant/alloantigen-related manuscripts. The IEDB curation involves an unprecedented level of detail, capturing for each paper the actual experiments performed for each different epitope structure. Key to enabling this process was the extensive use of ontologies to ensure rigorous and consistent data representation as well as interoperability with other bioinformatics resources, including the Protein Data Bank, Chemical Entities of Biological Interest, and the NIAID Bioinformatics Resource Centers. A growing fraction of the IEDB data derives from direct submissions by research groups engaged in epitope discovery, and is being facilitated by the implementation of novel data submission tools. The present explosion of information contained in biological databases demands effective query and display capabilities to optimize the user experience. Accordingly, the development of original ways to query the database, on the basis of ontologically driven hierarchical trees, and display of epitope data in aggregate in a biologically intuitive yet rigorous fashion is now at the forefront of the IEDB efforts. We also highlight advances made in the realm of epitope analysis and predictive tools available in the IEDB.

Keywords: B cells; MHC/HLA; T cells.

Introduction

The Immune Epitope Database and Analysis Resource[1] was created in 2003 with funding from the National Institute of Allergy and Infectious Disease (NIAID) to provide the scientific community with a central repository of freely accessible epitope data and an epitope prediction and analysis resource. Specifically, as stated by the original NIAID scope, the primary purpose of the project is to design, develop, populate and maintain a publicly accessible, comprehensive immune epitope database containing linear and conformational antibody epitopes and T-cell epitopes composed of MHC-binding peptides and ligands with a priority for epitopes associated with NIAID category A–C potential bioterrorism pathogens and their toxins (listed at http://www.niaid.nih.gov/dmid/biodefense/bandc_priority.htm).

In conjunction with the database component, an analysis resource has been developed that includes online access to: (i) tools to help researchers locate and analyse information contained in the Immune Epitope Database (IEDB); (ii) other relevant databases and related information; (iii) data-mining algorithms, mathematical models and other analytical tools to help researchers identify novel antibody and T-cell epitopes from genome or protein sequence information, predict the immunogenicity or antigenicity of epitopes, and predict host immune responses to particular epitopes. Herein, we highlight the progress of the IEDB from its inception to date, with particular focus on the three major areas – establishing and maintaining the database, establishing and maintaining the analysis resource, and maintaining close ties to the scientific community.

Establishing and maintaining the IEDB

Database design and IEDB ontology

In terms of design, one element that makes the IEDB unique among epitope databases is its association with a formalized ontology. The data housed in the IEDB is far more detailed than simple lists of epitope sequences. Rather, experimental details associated with each epitope, as reported in the literature, are also represented. Such experimental details (e.g. host organism, cell/antibody types, assays used) can encompass as many as 300 individual data fields, and therefore the IEDB demands a formal ontology to accurately and consistently represent these experimental details.

The development of the first rudimentary version of the IEDB ontology predated the actual curation of data, and therefore served as a blueprint for the initial IEDB design.[2] However, once the IEDB was established and significant quantities of data became available, a more refined and inclusive ontology was implemented (http://ontology.iedb.org), in conjunction with the Ontology of Biomedical Investigations (http://obi-ontology.org).[3] This pivotal undertaking guides our curation efforts, and is essential to rigorous and consistent data curation. Furthermore, this formal ontology has enabled us to represent complex immunological data in a streamlined, computer-readable format, hence facilitating data consistency validation, highly specific query formulation by IEDB users, and enhanced interoperability with other bioinformatics resources.[4] The formalized ontology developed to represent such complex immunological data are available in the ONTology of Immune Epitopes (ONTIE).[5]

Identification and categorization of relevant publications

The processes implemented by the IEDB team to facilitate curation of the large amount of highly detailed data relies heavily on a rigorous process used to identify the journal articles for curation, based on the combined use of PubMed queries, automated text classifiers and categorizers, and manual inspection of records by senior immunologists. First, the PubMed database (over 22 million papers) is queried to identify abstracts of potentially curatable manuscripts. Using this query, we have selected abstracts of potential relevance, spanning from the start of the scientific literature included in PubMed to the present date. Following this step, each abstract is further analysed to determine whether the reference is indeed curatable. The specific criteria are described in more detail in previous publications.[6–8] Developing an effective approach to handle this step was a key element in the development of the IEDB curation strategy, because we needed to minimize human inspection of the large number of abstracts, but also maintain low and acceptable error rates. To put this in perspective, the IEDB team has processed, over the duration of the 6·5 years in which curation has been in play, over 160 000 abstracts, corresponding to about 100 abstracts each working day. This milestone achievement was accomplished by establishing an advanced strategy that employed a semi-automated process, in which iteratively trained document classifiers are used to eliminate papers with high probability of being uncuratable.[9]

Furthermore, we developed additional classifiers that further classify each publication by its general topic (e.g. infectious diseases, autoimmunity, allergy, transplantation, HIV or cancer). This automated process also assigns the manuscript to a more specific subcategory. For example, an 'autoimmunity' paper may be placed in a 'diabetes' or 'multiple sclerosis' subcategory.[9,10] Manuscripts relating to a specific category can be processed as a group, allowing for resolution of category-specific issues while enhancing consistency. Most importantly, this clearly identifies papers from high priority subject areas (e.g. A–C pathogens), while also identifying manuscripts that relate to a category that is either a low IEDB priority, or out of the IEDB scope altogether. Senior immunologists manually inspect the curatable abstracts and the category into which they have been classified. The results are fed back into the classifier for retraining of the algorithm. This strategy has yielded a progressive increase in the accuracy and throughput of the automated screening.[8–10]

The categorization process has also revealed important trends with respect to global disease morbidity and mortality data. As we have reported, our classifications have shown a direct correlation between diseases associated with high morbidity and mortality and the degree to which they have been studied, while a relative scarcity of

Table 1.

Category	PubMed Query	Uncuratable by abstract scan	Potentially relevant after abstract scan	Curated	Uncuratable	Unavailable	% Done
Infectious Diseases	35 688	25 194	10 494	7484	2614	295	99·0
Allergy	7651	5962	1689	1178	402	82	98·4
Autoimmunity	14 936	9662	5274	3839	1203	161	98·7
Transplantation/Alloantigens	13 444	12 452	992	708	246	29	99·1
Others	89 921	84 637	5284	419	78	1467	37·2
Total	161 640	137 907	23 733	13 628	4543	2034	85·1
Total excluding 'Others'	71 719	53 270	18 449	13 209	4465	567	98·9

data exists for diseases of lesser global significance. Hence, a potentially powerful secondary outcome of our efforts can be to guide research efforts towards under-studied disease categories.[10]

Populating the IEDB

Following categorization, experimental data are curated by a team of doctorate-level curators using an optimized process and clearly defined guidelines, established with the expertise of senior immunologists.[11] As of April 2012, over 13 000 manuscripts have been selected as being within the scope of interest of the IEDB and have been curated (Fig. 1). For tracking and reporting purposes, the manuscripts were placed into five broad categories – infectious diseases, allergy, autoimmunity, transplantation/alloantigens and others (which are not a priority under the contract scope). Table 1 shows that all priority categories are near or have exceeded the 99% completion level. Hence, curation is in maintenance mode for all categories, representing a major accomplishment of this programme.

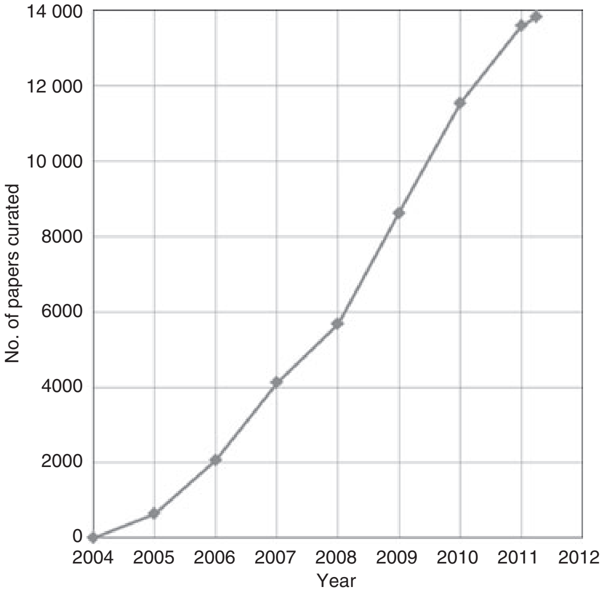


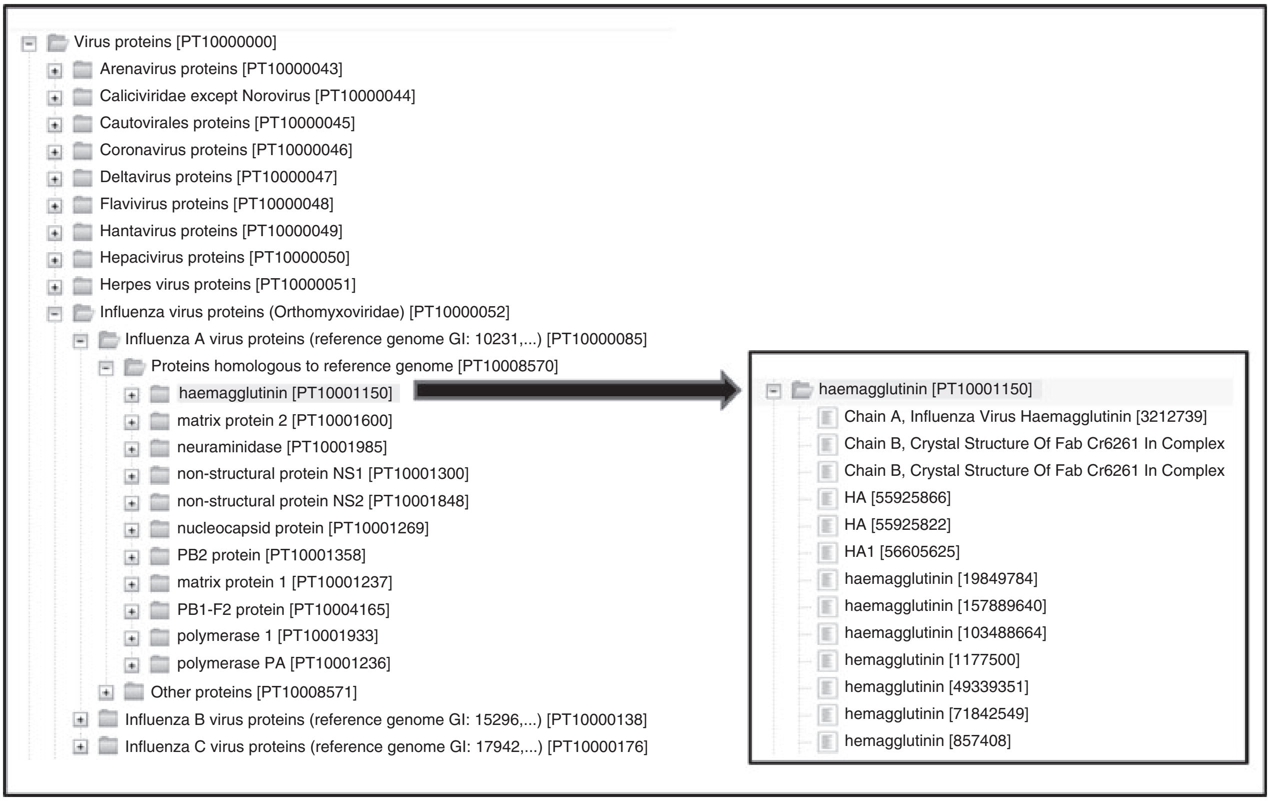
Figure 1. Total papers curated as of 1 April 2012.

The significant growth of the data available is also largely attributable to our progress in the curation of non-peptidic[12] epitopes, papers containing structural data defining epitopes bound to antibody, T-cell receptors or MHC molecules, as well as the inclusion of data directly submitted to the IEDB by scientists. To the best of our knowledge, non-peptidic molecules have not been curated in epitope databases until now. In addition, we have developed original processes to allow direct epitope submissions from immunological investigators, with emphasis on simplicity and web-accessibility. Just over 40 000 records encompassing about 26% of the IEDB are derived from direct submissions, attesting to the success of this endeavour. Collectively, the efforts to populate the IEDB during its formative years have resulted in the availability of over 88 000 epitopes, equivalent to the inclusion of about 40 epitopes per day.

Optimizing the query interface

The IEDB team has developed different query tools and strategies that reflect the varied needs of its composite user base. These search strategies have constantly been optimized based on feedback from the user community, and as the scope, type and size of the data housed in the IEDB has evolved.[1] To facilitate use of the main query interface on the IEDB home page and the advanced queries, the team iteratively developed a number of finders for selecting molecules, assays, source organisms, alleles and diseases. In so doing, we have taken advantage of the availability of the formal ontology described above, which provides the capacity of linking to other bioinformatics resources, thereby adding value to the IEDB data.

A prime example of such interoperability is demonstrated by the IEDB epitope source molecule finder. As the IEDB data set expanded, the inadequacies of the molecule finder became apparent. To select a specific source molecule, the user was forced to type in its name in a free text field. Although this is a workable method, it is error-prone and can result in unintended data omission. We



- Virus proteins [PT10000000]
  - Arenavirus proteins [PT10000043]
  - Caliciviridae except Norovirus [PT10000044]
  - Cautovirales proteins [PT10000045]
  - Coronavirus proteins [PT10000046]
  - Deltavirus proteins [PT10000047]
  - Flavivirus proteins [PT10000048]
  - Hantavirus proteins [PT10000049]
  - Hepacivirus proteins [PT10000050]
  - Herpes virus proteins [PT10000051]
  - Influenza virus proteins (Orthomyxoviridae) [PT10000052]
    - Influenza A virus proteins (reference genome GI: 10231,...) [PT10000085]
      - Proteins homologous to reference genome [PT10008570]
        - haemagglutinin [PT10001150]
        - matrix protein 2 [PT10001600]
        - neuraminidase [PT10001985]
        - non-structural protein NS1 [PT10001300]
        - non-structural protein NS2 [PT10001848]
        - nucleocapsid protein [PT10001269]
        - PB2 protein [PT10001358]
        - matrix protein 1 [PT10001237]
        - PB1-F2 protein [PT10004165]
        - polymerase 1 [PT10001933]
        - polymerase PA [PT10001236]
      - Other proteins [PT10008571]
    - Influenza B virus proteins (reference genome GI: 15296,...) [PT10000138]
    - Influenza C virus proteins (reference genome GI: 17942,...) [PT10000176]

- haemagglutinin [PT10001150]
  - Chain A, Influenza Virus Haemagglutinin [3212739]
  - Chain B, Crystal Structure Of Fab Cr6261 In Complex
  - Chain B, Crystal Structure Of Fab Cr6261 In Complex
  - HA [55925866]
  - HA [55925822]
  - HA1 [56605625]
  - haemagglutinin [19849784]
  - haemagglutinin [157889640]
  - haemagglutinin [103488664]
  - hemagglutinin [1177500]
  - hemagglutinin [49339351]
  - hemagglutinin [71842549]
  - hemagglutinin [857408]

Figure 2. The protein finder. Shown on the left is an excerpt of the hierarchical tree in which proteins are organized. The inset panel on the right shows protein entries from GenBank, the Protein Data Bank and UniProt that are categorized under the haemagglutinin node.

sought to devise a finder that would provide the ability to browse the available source molecules in a meaningful way, while also enabling the user to select a source protein and have all related proteins added to the query as well.

To this end, the National Center for Biotechnology Information (NCBI) source species was determined for each of the proteins in the database. For each source species, a set of reference proteins was selected from the NCBI protein database based upon the availability of a complete genome for the species, and proteins for each species were BLASTed against the reference protein set to determine their homologues. The result is a coherent tree that is divided along major taxonomic categories and is quickly traversed with proteins grouped logically below each species. It is now possible, for example, to select all Influenza A haemagglutinin proteins by selecting one node of the tree rather than individually clicking on the more than 100 different haemagglutinin proteins in the database (Fig. 2). The development of the protein tree has greatly simplified the process of browsing for a protein of interest, and represents a novel query mechanism that enhances access to the data in a biologically intuitive way. Similar linkages are being used with the NIAID-funded Bioinformatics Resource Centers[13–17] that maintain detailed information about the function, expression, structure and other features of proteins in infectious agents, as well as with the Protein Data Bank (PDB)[18] and Chemical Entities of Biological Interest (ChEBI) (http://www.ebi.ac.uk/chebi/).[19]

Establishing and maintaining the analysis resource

MHC class I binding predictions

Since the launch of the IEDB Analysis Resource (IEDB-AR), several new epitope prediction tools have been developed, while existing tools have been enhanced (Table 2). The growth over time in the availability of MHC class I peptide-binding data has led to the retraining and testing of all MHC class I peptide prediction methods. The correlation between quantity of available training data and prediction accuracy is evident in the improved prediction performances of the newly trained algorithms for MHC alleles for which new data sets were available.[20–24] In addition, the breadth of prediction coverage, in terms of different MHC molecules, also improved, including new alleles from humans, mice and non-human primates. In fact, the number of alleles for humans became so numerous that the option to limit the selectable alleles to those that occur in at least 1% of the human population was added.

Table 2.

Tool Category	Method
T Cell Epitope - MHC Class I Binding Prediction	Artificial Neural Network (ANN)
	Stabilized Matrix Method (SMM)
	Average Relative Binding (ARB)
	SMM with a peptide:MHC binding energy covariance matrix (SMMPMBEC)
	Scoring matrices derived from combinatorial peptide libraries (Comblib Sidney 2008)
	Consensus
	NetMHCpan
T Cell Epitope - MHC Class II Binding Prediction	Consensus
	Stabilized Matrix Method Aligin (SMM algin)
	Combinatorial Library
	Sturniolo
	Average Relative Binding (ARB)
	NN Align
	NetMHCIIpan
T Cell Epitope – Processing Prediction	Artificial Neural Network (ANN)
	Stabilized Matrix Method (SMM)
	Average Relative Binding (ARB)
	SMM with a peptide: MHC binding energy covariance matrix (SMMPMBEC)
	Scoring matrices derived from combinatorial peptide libraries (Comblib Sidney 2008)
	NetMHCpan
	NetChop
	NetCTL
B Cell Linear Epitope Prediction	Chou & Fasman Beta-Turn
	Emini Surface Accessibility
	Karplus & Schulz Flexibility
	Kolaskar & Tongaonkar Antigenicity
	Parker Hydrophilicity
	BepiPred
B Cell Discontinuous & Linear Epitope Prediction	DiscoTope
	ElliPro
Analysis Tools	Population Coverage
	Epitope Conservancy Analysis
	Epitope Cluster Analysis
	Homology Mapping

To best appreciate the progress in terms of class I binding predictions, one can compare the class I binding predictions available in 2003, when the IEDB-AR was initiated, with those available today. In terms of accuracy of prediction, previous area under the curve (AUC) values achieved were on average 0·796,[20] whereas they are currently 0·896.[25] In terms of breadth, binding predictions are now available for hundreds of MHC class I alleles encompassing humans, non-human primates, mice and other species. Notably, the suite of algorithms provides predictions for over 75 of the most common HLA A and B specificities in the worldwide population, whereas predictions were available for approximately 40 alleles 6 years ago.

MHC class II binding predictions

Unlike the closed binding groove of MHC class I molecules, which limits the size of the peptide that can be accommodated, the MHC class II peptide binding groove is open at both ends, enabling it to bind peptides of variable length.[26] The variability in peptide binding length represents a complicating factor in MHC class II binding prediction,[20,27,28] and results in an overall lower prediction accuracy compared with MHC class I predictions. Multiple MHC class II binding prediction services are hosted at the IEDB-AR.[21,29,30] Three new approaches were recently added, namely a combinatorial peptide library-based approach, which is generated independently of individual peptide-binding data, and NN-align and NetMHCIIpan, which were developed at the Center for Biological Sequence Analysis at the Technical University of Denmark and are based on a neural network approach.[31–33]

Again, comparisons between IEDB class II binding predictions available in 2003 with those available today reveal significant progress made in this realm. In terms of accuracy, previous AUC values were on average 0·67,[34] whereas current AUCs achieve values as high as 0·88. In terms of breadth, binding predictions are now available for hundreds of MHC class II alleles, compared with a dozen alleles initially. With the recent inclusion of the most common DP and DQ molecules, the suite of algo-

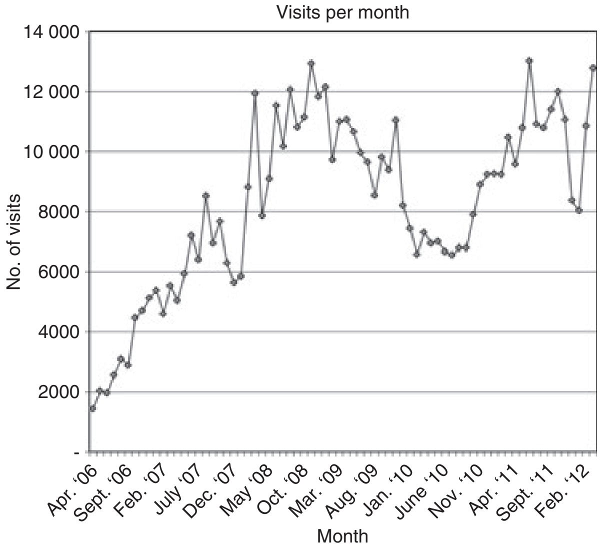


Figure 3. Number of visits to the Immune Epitope Database website per month from 1 April 2006 through to 1 April 2012.

rithms now provides predictions for the most common HLA DP, DQ and DR specificities in the worldwide population, allowing coverage of the general population in excess of 80% at each individual locus.

Prediction of B-cell epitopes

The IEDB-AR includes three newly developed methods for B-cell epitope predictions. First, BepiPred combines a position-specific scoring matrix with a propensity scale method to make linear epitope predictions.[35] DiscoTope, on the other hand, is a method for discontinuous epitope prediction using protein three-dimensional structural data as input.[36] As such, it performs better predictions of discontinuous epitopes than methods based exclusively on linear amino acid sequence information.[36] Both BepiPred and DiscoTope have been implemented with a new feature that enables the predicted discontinuous epitopes to be visualized on the three-dimensional structural image of the protein. Finally, the ElliPro (derived from Ellipsoid and Protrusion) method is based on a prediction strategy originally proposed by Thornton *et al.*,[37] who found a correspondence between regions of high protrusion index values and continuous epitopes defined experimentally using myoglobin, lysozyme and myohaemerythrin as models.[37] ElliPro combines this strategy with the MODELLER program,[38] a residue-clustering algorithm, and the Jmol viewer to predict and visualize antibody epitopes found in protein sequences and structures.

Maintaining close ties to the scientific community

Ultimately, the success of such a resource as the IEDB can be measured in terms of its use by the scientific community. Outreach activities provide a unique opportunity to promote awareness and use of the IEDB, while also gathering feedback to guide improvements and refinements. For these reasons, the IEDB team has emphasized organized outreach as a key component of the IEDB project. In the following paragraphs, we briefly review the impact of these outreach activities, as judged by criteria such as citations, website traffic, web links, interactions with other bioinformatics resources, and industrial licenses.

The IEDB staff has authored a total of 52 peer-reviewed articles or book chapters. IEDB publications have been well received by the scientific community and have received over 500 citations in scientific journals.[39] Furthermore, the IEDB staff has publicized its contents and mission in 65 talks at scientific meetings and lectures, and by 13 booths at various scientific meetings. Additionally as a critical component of our outreach strategy, we have used five scientific workshops specifically designed to gather feedback from community experts, through meta-analysis and related activities.

The IEDB also collects website usage statistics, which indicate a steadily growing user base. The number of visits per month is now about 11 000 in aggregate for the two servers hosting the main IEDB and Analysis Resource sites (Fig. 3). Furthermore, the number of unique visitors per month is about 3800 for the main website server and 1800 for the Analysis Resource server. The number of page views per month on the main website has grown to about 100 000 and the number of database export downloads varies between 40 and 80 per month. The latter statistic indicates that users are downloading the IEDB data into their local environment for further analysis and queries.

Another metric of community use is the web links that are embedded in websites or other documents on the internet. A total of 2010 distinct sites that contain links to the IEDB are found by searching Google for link:http: iedb.org. This includes links from related databases such as the Influenza Research Database (www.fludb.org), scientific papers available online, inclusion in link lists from professional societies such as the American Association of Immunologists (www.aai.org) or collections of tools such as bioinformatics.org.

Since 2009, the IEDB has offered a LinkOut protocol allowing other websites to establish links to their sites from relevant data on the IEDB website. Links were initially established with EuPathDB,[13] and then expanded to include the Influenza Research Database,[16] Pathosystems Resource Integration Center,[17] and Virus Pathogen Database and Analysis Resource.[15] We have also created links from all of the IEDB HLA molecules to their counterparts at the dbMHC database[40] hosted by the NCBI. In 2011, IEDB staff created links from the HLA alleles in the IEDB to the International Immunogenetics MHC Database.[41]

Consistent with its scope and mission, the IEDB is freely available worldwide, through web-based access. However, we have been approached by many of the largest pharmaceutical companies worldwide with requests to license the IEDB. The rationale for this request stems from the fact that these companies would prefer to use epitope analysis tools without disclosing over the internet the sequences of their vaccine or drug candidates being investigated. In line with the overall mission of leveraging the IEDB as a resource to freely advance vaccine and drug development, we have made these licenses available for a nominal fee intended to cover additional development and support effort. We are encouraged that large pharmaceutical companies recognize the value of the IEDB, and that the resource is being used by large vaccine and drug developers, as a tangible indication of its impact and usefulness. Our policy has been to make these licenses available free to non-profit research and educational organizations.

Conclusion

Over the past decade, the IEDB team has conceived, designed, built and deployed the most expansive and inclusive database and analysis resource of antibody and T-cell epitopes available to the public. Herein, we have detailed the major improvements made to the IEDB to date, both in terms of data content, and in terms of the user interface. At present, the IEDB data content is current in the categories of infectious disease, allergy, autoimmunity and transplant, and we will continue to curate newly published literature in all of these categories on a quarterly basis. As we embark upon a second funding cycle, we look forward to continuing to address the inherent challenge of offering features that facilitate more effective querying and interpretation of the rapidly growing dataset. We plan to continue to exploit the interactions with our user base and subject matter experts to evaluate and improve the IEDB interface and reporting schemes.

Acknowledgements

Team members include prime contractor La Jolla Institute for Allergy and Immunology and its subcontractors at Science Applications International Corporation, the San Diego Supercomputer Center at the University of California San Diego, the University of Copenhagen, the Technical University of Denmark, and the Chemical Entities of Biological Interest.

Disclosures

The authors have no conflicts of interest.

References

1 Vita R, Zarebski L, Greenbaum JA *et al.* The immune epitope database 2.0. *Nucleic Acids Res* 2010; **38**:D854–62.

2 Sathiamurthy M, Peters B, Bui HH *et al.* An ontology for immune epitopes: application to the design of a broad scope database of immune reactivities. *Immunome Res* 2005; **1**:2.

3 Brinkman RR, Courtot M, Derom D *et al.* Modeling biomedical experimental processes with OBI. *J Biomed Semantics* 2010; **1**(Suppl. 1):S7.

4 Peters B, Sette A. Integrating epitope data into the emerging web of biomedical knowledge resources. *Nat Rev Immunol* 2007; **7**:485–90.

5 Greenbaum JA, Kotturi MF, Kim Y *et al.* Pre-existing immunity against swine-origin H1N1 influenza viruses in the general human population. *Proc Natl Acad Sci U S A* 2009; **106**:20365–70.

6 Vita R, Peters B, Sette A. The curation guidelines of the immune epitope database and analysis resource. *Cytometry A* 2008; **73**:1066–70.

7 Salimi N, Vita R. The biocurator: connecting and enhancing scientific data. *PLoS Comput Biol* 2006; **2**:e125.

8 Wang P, Morgan AA, Zhang Q, Sette A, Peters B. Automating document classification for the Immune Epitope Database. *BMC Bioinformatics* 2007; **8**:269.

9 Seymour E, Damle R, Sette A, Peters B. Cost sensitive hierarchical document classification to triage PubMed abstracts for manual curation. *BMC Bioinformatics* 2011; **2**:482–92.

10 Davies V, Vaughan K, Damle R, Peters B, Sette A. Classification of the universe of immune epitope literature: representation and knowledge gaps. *PLoS ONE* 2009; **4**:e6948.

11 Vita R, Vaughan K, Zarebski L *et al.* Curation of complex, context-dependent immunological data. *BMC Bioinformatics* 2006; **7**:341.

12 Vita R, Peters B, Josephs Z *et al.* A Model for Collaborative Curation, The IEDB and ChEBI Curation of Non-peptidic Epitopes. *Immunome Res* 2011; **7**:1–8.

13 Aurrecoechea C, Brestelli J, Brunk BP *et al.* EuPathDB: a portal to eukaryotic pathogen databases. *Nucleic Acids Res* 2010;**38**:D415–9.

14 Lawson D, Arensburger P, Atkinson P *et al.* VectorBase: a data resource for invertebrate vector genomics. *Nucleic Acids Res* 2009; **37**:D583–7.

15 Pickett BE, Sadat EL, Zhang Y *et al.* ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res* 2012; **40**:D593–8.

16 Squires RB, Noronha J, Hunt V *et al.* Influenza Research Database: an integrated bioinformatics resource for influenza research and surveillance. *Influenza Other Respi Viruses* 2012. doi: 10.1111/j.1750-2659.2011.00331.x. [Epub ahead of print]

17 Gillespie JJ, Wattam AR, Cammer SA *et al.* PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect Immun* 2011; **79**:4286–98.

18 Rose PW, Beran B, Bi C *et al.* The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res* 2011; **39**:D392–401.

19 de Matos P, Alcántara R, Dekker A *et al.* Chemical Entities of Biological Interest: an update. *Nucleic Acids Res* 2010; **38**:D249–54.

20 Peters B, Bui H-H, Frankild S *et al.* A Community Resource Benchmarking Predictions of Peptide Binding to MHC-I Molecules. *PLoS Comput Biol* 2006; **2**:e65.

21 Bui H-H, Sidney J, Peters B *et al.* Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics* 2005; **57**:304–14.

22 Peters B, Sette A. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics* 2005; **6**:132.

23 Peters B, Tong W, Sidney J, Sette A, Weng Z. Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. *Bioinformatics* 2003; **19**:1765–72.

24 Nielsen M, Lundegaard C, Worning P, Lauemoller SL, Lamberth K, Buus S, Brunak S, Lund O. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci* 2003; **12**:1007–17.

25 Kim Y, Sidney J, Pinilla C, Sette A, Peters B. Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. *BMC Bioinformatics* 2009; **10**:394.

26 Jones EY, Fugger L, Strominger JL, Siebold C. MHC class II proteins and disease: a structural perspective. *Nat Rev Immunol* 2006; **6**:271–82.

27 Gowthaman U, Agrewala JN. *In silico* tools for predicting peptides binding to HLA-Class II molecules: more confusion than conclusion. *J Proteome Res* 2008; **7**:154–63.

28 Lin HH, Ray S, Tongchusak S, Reinherz EL, Brusic V. Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research. *BMC Immunology* 2008; **9**:8.

29 Nielsen M, Lundegaard C, Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics* 2007; **8**:238.

30 Sturniolo T, Bono E, Ding J *et al.* Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol* 1999; **17**:555–61.

31 Wang P, Sidney J, Kim Y, Sette A, Lund O, Nielsen M, Peters B. Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC Bioinformatics* 2010; **11**:568.

32 Nielsen M, Lund O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics* 2009; **10**:296.

33 Nielsen M, Lundegaard C, Blicher T, Peters B, Sette A, Justesen S, Buus S, Lund O. Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. *PLoS Comput Biol* 2008; **4**:e1000107.

34 Wang P, Sidney J, Dow C, Mothe B, Sette A, Peters B. A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS Comput Biol* 2008; **4**:e1000048.

35 Larsen J, Lund O, Nielsen M. Improved method for predicting linear B-cell epitopes. *Immunome Res* 2006; **2**:2.

36 Andersen PH, Nielsen M, Lund O. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci* 2006; **15**:2558–67.

37 Thornton JM, Edwards MS, Taylor WR, Barlow DJ. Location of 'continuous' antigenic determinants in the protruding regions of proteins. *EMBO J* 1986; **5**:409–13.

38 Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A. Comparative protein structure modeling using MODELLER. *Curr Protoc Protein Sci* 2007; **Chapter 2**:Unit 2 9.

39 Salimi N, Fleri W, Peters B, Sette A. Design and utilization of epitope-based databases and predictive tools. *Immunogenetics* 2010; **62**:185–96.

40 Sayers EW, Barrett T, Benson DA *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2011; **39**:D38–51.

41 Robinson J, Mistry K, McWilliam H, Lopez R, Parham P, Marsh SG. The IMGT/HLA database. *Nucleic Acids Res* 2011; **39**:D1171–6.