BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 14 2008, pages 1639–1640
doi:10.1093/bioinformatics/btn251

Sequence analysis

Web-based design and evaluation of T-cell vaccine candidates

James Thurmond[1,†], Hyejin Yoon[1,†], Carla Kuiken[1], Karina Yusim[1], Simon Perkins[2], James Theiler[1], Tanmoy Bhattacharya[1,3], Bette Korber[1,3] and Will Fischer[1,*]

[1]Los Alamos National Laboratory, Los Alamos, NM 87545, [2]UltraSpectral Inc., 5701 Carmel Ave. NE, Suite C, Albuquerque NM 87113 and [3]The Santa Fe Institute, Santa Fe, NM 87501, USA

Received on January 7, 2008; revised on May 8, 2008; accepted on May 28, 2008

Advance Access publication May 29, 2008

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: We present a suite of on-line tools to design candidate vaccine proteins, and to assess antigen potential, using coverage of *k*-mers (as proxies for potential T-cell epitopes) as a metric. The vaccine design tool uses the recently published 'mosaic' method to generate protein sequences optimized for coverage of high-frequency *k*-mers; the coverage-assessment tools facilitate coverage comparisons for any potential antigens. To demonstrate these tools, we designed mosaic protein sets for B-clade HIV-1 Gag, Pol and Nef, and compared them to antigens used in a recent human vaccine trial.

Availability: http://hiv.lanl.gov/content/sequence/MOSAIC/

Contact: wfischer@lanl.gov

Supplementary information: Supplementary data are available at ftp://ftp-t10.lanl.gov/pub/btk/WebToolsData

1 INTRODUCTION

Recent work on vaccines for highly variable pathogens (e.g. HIV-1) has focused on cell-mediated immunity (Weaver *et al.*, 2006), using synthetic antigens that include much of the sequence diversity of pathogen populations (Gaschen *et al.*, 2002; Nickle *et al.*, 2007). Such antigens (e.g. consensus-, ancestral- or center-of-tree sequences) are not natural proteins, but resemble them closely enough to be immunologically functional by various criteria (reviewed in Brander *et al.*, 2007).

Our mosaic vaccine method (Fischer *et al.*, 2007) generates sets of optimized synthetic proteins, each one a patchwork of natural protein subsequences. Mosaics are highly similar to intact full-length natural proteins (and hence likely to preserve processing and HLA presentation), but they are optimized for maximal coverage of short '*k*-mer' fragments (9–12 amino acids). Mosaics have been shown to be immunogenic (in mice, G. Nabel, personal communication; non-human primate trials are in progress); they exclude unnatural and rare *k*-mers (notably at assembly breakpoints, precluding the possibility of 'neoantigens'), and include the most common variants from large and possibly very diverse populations of natural sequences.

The HIV Sequence Database team at the Los Alamos National Laboratory has developed web-based tools for designing vaccine cocktails and for assessing the potential epitope coverage of any sequence set. The Mosaic Vaccine Designer Tool distills an input set of protein sequences into a much smaller set of mosaic protein sequences. The Vaccine Epitope Coverage Assessment Tool (Epicover) and the Positional Epitope Coverage Tool (Posicover) are adjunct tools that compute how well a proposed 'antigen set' covers potential epitopes in a 'test-set' of natural sequences. An antigen set could be a mosaic cocktail, other potential vaccine proteins or a set of peptide reagents for assessing T-cell responses. *Epicover* calculates the mean coverage of the test-set population by antigen-set *k*-mers; *Posicover* provides detailed positional coverage information relative to a test-set alignment. Both tools provide graphical output and allow detailed user control.

2 METHODS

2.1 Mosaic Vaccine Designer

Input protein sequences (the 'training set') can be provided in most common sequence formats (alignment not needed). In 'basic' mode, users choose the major parameters of the genetic algorithm: e.g. the number of sequences to be included in the final mosaic cocktail, length of potential epitopes [the default value is 9, for CD8+ T-cell epitopes (Marsh *et al.*, 2000, p. 69)]. Alternatively, one or more mosaic sequences may be generated as 'add-ons' to complement one or more fixed sequences (e.g. natural or consensus proteins). A 'rarity threshold' excludes very low-frequency *k*-mers from the mosaics. This threshold can be adjusted for training sets of varying size or diversity. The algorithm performs a series of replicate runs, each using a different random starting set, halting when the rate of coverage score increase drops below a threshold (the stopping criterion). For comparison, or as an alternative vaccine strategy, the best set of *n* natural proteins (in terms of potential epitope coverage) can also be selected from the input sequences. An 'advanced' option in the interface provides more detailed control of the algorithm. Adjustable parameters include stopping criterion, crossover probabilities and maximum runtime.

2.2 Coverage assessment tools (Epicover, Posicover)

Two tools are available to compare the diversity coverage of vaccine antigens (such as the mosaic sets created by the Mosaic Vaccine Designer) or peptide reagents (e.g. for T-cell response assays). Both tools compute coverage of one or more user-specified test sets by one or more antigen sets. Users can specify the nominal epitope length (nine amino acids by default).

Potential epitope coverage is calculated using the optimization metric used by the Mosaic Vaccine Designer tool, except that rare or unique *k*-mers (not scored by the Mosaic Vaccine Designer) are scored if present. Because similar epitopes may cross-react, both exact-match and near-match

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

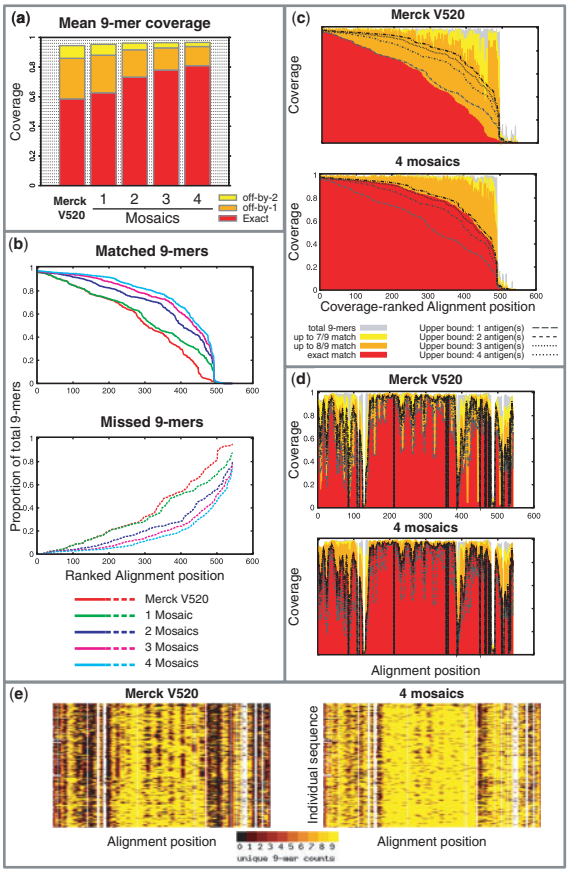

Fig. 1. Coverage plots for two vaccine candidates. Mosaic sequences were generated, and antigen sequences assessed, using a B-clade Gag alignment. (**a**) Epicover: coverage of test-set by five 'antigen-sets' (Merck V520 trial antigens and four mosaic sets). (**b–e**) Posicover: (b) 9mers in the test set matched or missed by five vaccine sets. (c–e) Vaccine trial antigens versus 4-sequence mosaic set. (c) Ranked 9mer coverage. (d) Coverage by column in the aligned test-set. (e) Coverage mapped on the input alignment, amino acid represented by single pixels, color-coded by the number of unique antigen 9mers covering it (black denotes 'missed' *k*-mers; yellow, complete coverage). (b,c) are sorted by coverage score.

scores are computed. Publication-quality plots can be downloaded in various formats (e.g. EPS, PDF, PNG).

2.2.1 Epicover The epitope coverage assessment tool computes coverage values for all antigen-set/test-set combinations. The fraction of *k*-mers shared with the antigen set is calculated for each test-set sequence; the per-sequence mean is reported. Test-set subsets (e.g. clades or geographic regions) can be defined and scored individually. Epicover also reports counts of antigen-set *k*-mers that are rare, unique or absent from the test set.

2.2.2 Posicover The positional epitope coverage assessment tool presents coverage by position in a sequence alignment. The various plots (Fig. 1b–e) show the proportion of covered *k*-mers (present in both the test-set and the antigen-set), including partial matches, and of missed *k*-mers (present in the test set but not the antigen set), for all alignment positions. Ranked coverage plots (Fig. 1b, d) simplify comparisons between sets; plots ordered by alignment position (Fig. 1c, e) show coverage differences between regions of the target protein. A final plot (Fig. 1e) shows potential epitope coverage on a sequence alignment. Local *k*-mer coverage is clearly revealed by color coding each amino acid by the number of antigen-set *k*-mers that include it.

3 CONCLUSIONS

Mosaic antigens 'cover' potential epitopes better than the antigens from the Merck V520 trial; this implies a greater likelihood of inducing protective responses against diverse HIV-1 strains (Fischer *et al.*, 2007). High population coverage is achieved with only three sequences (Fig. 1a, b), suggesting that mosaic vaccine antigens, if immunologically effective, may be economically practical as well.

Immune responses involve many factors besides sequence identity between infectious agent and vaccine. The mosaic method optimizes only sequence identity; the vaccine candidates it produces will require experimental evaluation in varied immunological backgrounds. It is striking, however, that the Fischer *et al.* (2007) mosaics contained the majority of 'elite' CD8+ T-cell epitopes later identified in an algorithmic/immunological study (Perez *et al.*, 2008).

These tools simplify generation and quantitative evaluation of prospective antigens, and could be applied to any variable pathogen.

ACKNOWLEDGEMENTS

The authors thank Andrew Bett and Michael Robertson for the Merck V520 sequences and Paul Fenimore for useful comments.

Funding: This work was funded by Los Alamos National Laboratory (LDRD-DR); NIH (HIV-RAD PO1 AI61734, to B.K. at the Santa Fe Institute).

Conflict of Interest: none declared.

REFERENCES

Brander,C. *et al.* (2007) Capturing viral diversity for in-vitro test reagents and HIV vaccine immunogen design. *Curr. Opin. HIV AIDS*, **2**, 183–188.

Fischer,W. *et al.* (2007) Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. *Nat. Med.*, **13**, 100–106.

Gaschen,B. *et al.* (2002) Diversity considerations in HIV-1 vaccine selection, *Science*, **296**, 2354–2360.

Marsh,S. *et al.* (2000) *The HLA FactsBook*. Academic Press, London, San Diego.

Nickle,D. *et al.* (2007) Coping with viral diversity in HIV vaccine design, *PLoS Comput. Biol.*, **3**, e75.

Pérez, C.L. *et al.* (2008) Broadly immunogenic HLA Class I supertype-restricted elite CTL epitopes recognized in a diverse population infected with different HIV-1 subtypes. *J. Immunol.*, **180**, 5092–5100.

Weaver,E.A. *et al.* (2006) Cross-subtype T-cell immune responses induced by a Human Immunodeficiency Virus Type 1 Group M consensus Env immunogen. *J. Virol.*, **80**, 6745–6756.