Regulatory Toxicology and Pharmacology 67 (2013) 232–239

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

Comparative assessment of multiple criteria for the *in silico* prediction of cross-reactivity of proteins to known allergens

Henry P. Mirsky *, Robert F. Cressman Jr. [1], Gregory S. Ladics [1]

DuPont Pioneer, Route 141, Henry Clay Bldg, #400, Wilmington, DE 19880-0400, USA

ARTICLE INFO

Article history:
Received 4 February 2013
Available online 9 August 2013

Keywords:
Allergenic cross-reactivity
In silico
Bioinformatics
Assessment criteria
False positives
False negatives
Sensitivity
Specificity

ABSTRACT

Genetically modified crops are becoming important components of a sustainable food supply and must be brought to market efficiently while also safeguarding the public from cross-reactivity of novel proteins to known allergens. Bioinformatic assessments can help to identify proteins warranting further experimental checks for cross-reactivity. This study is a large-scale *in silico* evaluation of assessment criteria, including searches for: alignments between a query and an allergen having ⩾35% identity over a length ⩾80; any sequence (of some minimum length) found in both a query and an allergen; any alignment between a query and an allergen with an *E*-value below some threshold. The criteria and an allergen database (AllergenOnline) are used to assess 27,243 *Viridiplantae* proteins for potential allergenicity. (A protein is classed as a "real allergen" if it exceeds a test-specific level of identity to an AllergenOnline entry; assessment of real allergens in the query set is against a reduced database from which the identifying allergen has been removed.) Each criterion's ability to minimize false positives without increasing false negative levels of current methods is determined. At best, the data show a reduction in false positives to ~6% (from ~10% under current methods) without any increase in false negatives.

1. Introduction

Prior to commercialization, genetically modified (GM) crops undergo an extensive safety evaluation. One component of this evaluation is the assessment of the newly expressed protein(s) for allergenic potential. Currently, no single factor is recognized as an identifier for protein allergenicity. However, *in silico* screening of genetic modifications has become mandatory and may help to identify novel proteins possessing possible allergenic cross-reactivity – and therefore warranting subsequent experimental tests for confirmation (e.g., IgE-binding studies, pepsin digestibility studies, etc.). These data (and their quality) are then weighed to reach a conclusion regarding risk (Codex Alimentarius Commission, 2009).

Near-identity of candidate protein sequences with known allergens will certainly identify them as such, if that is the case, and serum testing would presumably be unnecessary. For those query sequences that do not clearly match known allergens, bioinformatic assessment, although not a replacement for empirical assessment of IgE binding, can provide insight into possible cross-reactivity, as sequences sharing a high degree of identity often share immunologically relevant topology (Aalberse, 2000; Aalberse et al., 2001; Goodman et al., 2008.) Traditionally, such analysis consists of two components: A search for short linear epitopes as well as comparison of primary amino acid sequences using FASTA (Pearson and Lipman, 1988) or BLAST (Altschul et al., 1997) to locate possible shared conformations (Cressman and Ladics, 2009). Recent publications (EFSA, 2011; Goodman, 2008; Goodman and Tetteh, 2011; Harper et al., 2012; Herman et al., 2009; Ladics et al., 2011; Young et al., 2012) have argued that the standard search for a sequence of eight or more amino acids found in both the query and a known allergen provides little value; however it has been demonstrated that insertion of a short stretch of amino acids derived from a known allergen into the correct conformational context can result in an increase in specific IgE binding (Klinglmayr et al., 2009). Additionally, the default local alignment search criteria have been constrained by the imposition of a defined threshold (⩾35% sequence identity over an alignment length ⩾80) (Ladics et al., 2007). This constraint neglects many of the features that help to define relevant homologies between sequences, features incorporated into the algorithms themselves (e.g., *E*-value [expectation value], which is a measure of the relatedness between protein sequences). A small *E*-value (e.g., 10^{-8}) indicates a potential biologically relevant similarity in the context of potential allergenic cross-reactivity; large *E*-values (>1.0) represent random alignments that do not possess biologically relevant similarity (Pearson, 2000; Silvanovich et al., 2009).

* Corresponding author. Fax: +1 302 695 3075.
E-mail addresses: henry.mirsky@pioneer.com (H.P. Mirsky), robert.f.cressman@pioneer.com (R.F. Cressman Jr.), gregory.s.ladics@usa.dupont.com (G.S. Ladics).
[1] Fax: +1 302 695 3075.

0273-2300/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.yrtph.2013.08.001

In order to investigate the refinement of current bioinformatic practices for assessing potential cross-reactivity, we have undertaken a systematic investigation of the effect of various *in silico* assessment criteria on the ability to distinguish real allergens from non-allergens. The various criteria include one or more of the following components: (1) A check for the presence of an alignment between a query protein and an allergen having ⩾35% sequence identity over an alignment length ⩾80 (Codex Alimentarius Commission, 2009); (2) a check for the presence of a sequence, at or above a specified length, found in both a query protein and an allergen; and (3) a check for the presence of an alignment between a query protein and an allergen with an *E*-value at or below some specified threshold. While there have been previous investigations into the way some of these criteria affect detection rates (Silvanovich et al., 2009), to our knowledge there have been no studies on this scale. Our analyses have uncovered numerous potential allergenic assessment criteria that are able to lower the current percent of false positives without raising the current percent of false negatives – at best, a reduction in the percent of false positives from ~10% to ~6% (a >40% reduction) is achieved without any increase in the percent of false negatives and with only minor changes to the existing criteria.

To date, there have been no known instances of transgenic proteins in GM crops inducing responses in humans. As food demand increases while environmental pressures grow, it is essential that the adoption of promising crop traits not be delayed by extraneous experimental investigations caused by positive but misleading bioinformatic matches. Therefore, *in silico* predictions of possible cross-reactivity should be as useful as possible, and it is hoped that the results provided here will help to serve that end.

2. Methods and materials

2.1. Query proteins

The entire set (31,897 sequences) of reviewed *Viridiplantae* proteins was obtained from UniProtKB/Swiss-Prot (http://www.uniprot.org). Under current Codex guidelines, an alignment of a protein to an allergen must possess a length ⩾80; therefore, sequences with lengths smaller than 80 amino acids were removed from further consideration. Additionally, any sequence annotated as a "fragment" was removed. The final query protein dataset contained 27,243 protein sequences. See Fig. 1.

2.2. Allergens

Version 12 (January 2012) of the AllergenOnline database (http://www.allergenonline.org), a peer-reviewed allergen database compiled by the Food Allergy Research and Resource Program (FARRP) (http://farrp.unl.edu) at the University of Nebraska, Lincoln, was used for analyses. AllergenOnline v12 contains 1603 protein sequences.

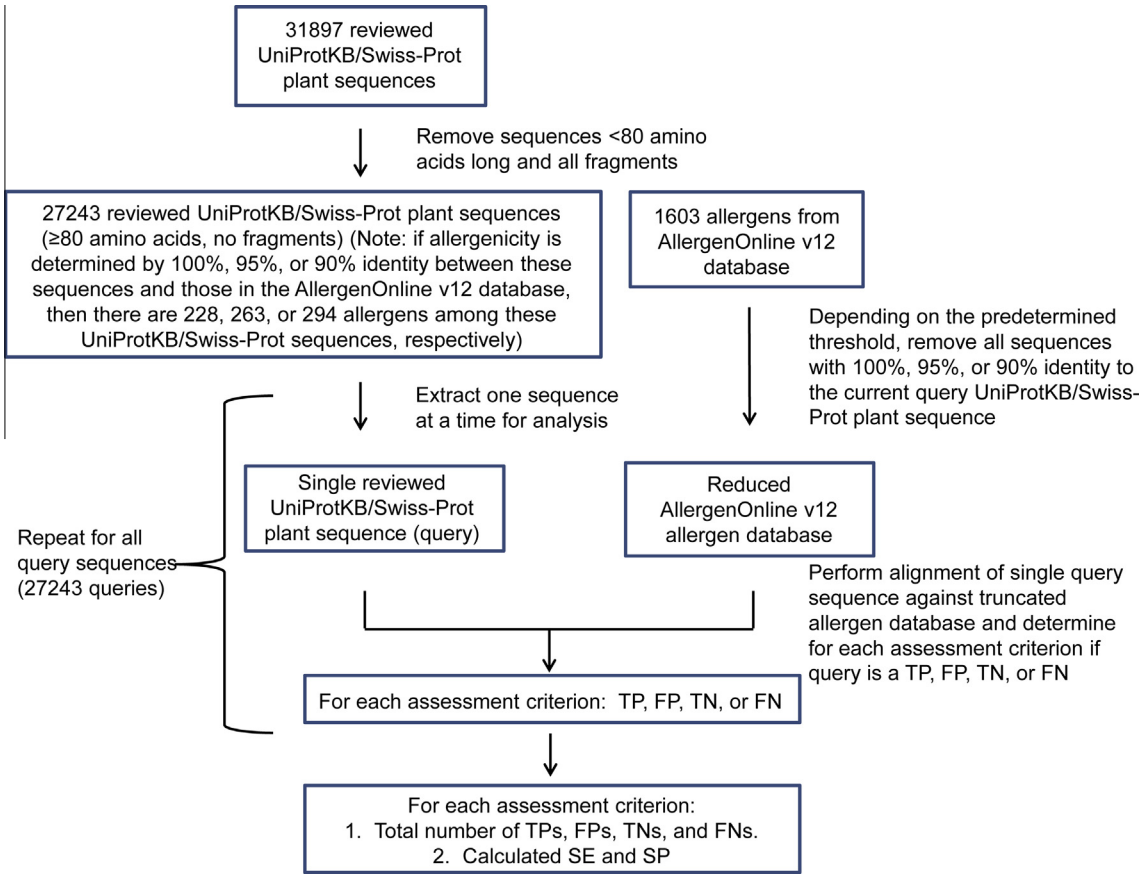

Fig. 1. Process flow chart. The entire set of reviewed *Viviplantae* proteins (31,897 proteins) was downloaded from UniProtKB/Swiss-Prot (http://www.uniprot.org). All proteins smaller than 80 amino acids in length, as well as all fragments, were removed. The first of the remaining 27,243 proteins was selected. This protein was compared against the 1603 proteins in the Food Allergy Research and Resource Program's (FARRP) AllergenOnline (http://www.allergenonline.org) (Version 12) (AllergenOnline v12) dataset and declared a "*real allergen*" if sequence identity was 100%, ⩾95%, or ⩾90%, depending on the test. The matching sequences in AllergenOnline v12, if any, were removed from AllergenOnline v12 to form a reduced allergen database. The query protein was evaluated for allergenicity against the reduced allergen database for each specified allergenic assessment criterion and declared either a true positive (TP), false positive (FP), true negative (TN), or false negative (FN). The original AllergenOnline v12 database was restored, and the process was repeated for each of the remaining query proteins. The total number of TPs, FPs, TNs, and FNs produced using each allergenic assessment criterion was used to calculate the criterion's sensitivity (SE) and specificity (SP).