DOI: 10.1478/C1A1002006

AAPP | Atti della Accademia Peloritana dei Pericolanti
Classe di Scienze Fisiche, Matematiche e Naturali
ISSN 1825-1242

Vol. LXXXVIII, No. 2, C1A1002006 (2010)

IN SILICO ALLERGEN IDENTIFICATION: PROPOSAL FOR A REVISION OF FAO/WHO GUIDELINES

FABRIZIO GUARNERI [a]

(Communication presented by Prof. Giacomo Tripodi)

ABSTRACT. Allergy is a widespread, often severe health problem. *In vivo* or *in vitro* identification of new allergenic proteins (natural or bioengineered) is time- and resource-consuming, and *in vivo* testing can be dangerous. Thus, allergenicity prediction through computation (*in silico*) was proposed to narrow down the number of potential allergens to be tested with traditional methods. In 2001, the Food and Agriculture Organization (FAO) and the World Health Organization (WHO) officially defined guidelines for *in silico* allergenicity prediction, based on amino acid sequence similarity to known allergens; these guidelines, however, have been criticized because of frequent false positives. In the present work, the BLAST (Basic Local Alignment Search Tool) software was used to compare known and potential allergens, and select only statistically significant homologies (i.e. homologies whose E value, calculated by BLAST, was < 1); FAO/WHO rules were then applied to these homologies. With this method, correct recognition of all known allergens, with only 10 false positives (1.26% of all predicted allergens) was achieved when using an upper limit of 0.1 for E values; complete suppression of wrong predictions, while maintaining 100% sensitivity, was obtained with little modifications of the minimum requirements contained in the FAO/WHO guidelines.

1. Background

Allergy is a common and widespread health problem, particularly frequent in urban areas of industrialized countries. It is medically defined as the hypersensitivity/hyperreactivity of the immune system against harmless environmental molecules. From an immunological point of view, two types of allergy can be distinguished:

- immediate hypersensitivity or IgE-mediated (mediated by class E immunoglobulins) hypersensitivity, also defined "type I reaction" in the Gell and Coombs' classification of immune reactions [1];
- delayed-type hypersensitivity or cell-mediated hypersensitivity, defined as "type IV reaction" in the Gell and Coombs' classification [1].

The above reactions are classically elicited by different antigens (i.e. molecules against which an immune response can be mounted). Type I reactions, for which the term "allergy" is more often used, are caused by allergens, i.e. antigens that can induce the production of, and are bound by, specific IgE antibodies. Typically, these reactions are involved in

several cases of rhinitis, asthma, conjunctivitis, urticaria/angioedema, anaphylactic shock. In the case of type IV reactions, instead, a hapten, i.e. a small molecule that is per se unable to elicit immune response, binds to a host molecule and chemically modifies it, thus generating a complete antigen. Allergic contact dermatitis is the most common clinical manifestation of this kind of hypersensitivity.

Although much has been done in the research on allergens, some problems are still open:

- identification of new allergens;
- identification of cross-reactivity between allergens (persons sensitized to a substance can react to a different, but structurally similar substance; this can be particularly annoying or even dangerous when the cross-reactivity is not known);
- definition of the allergenic profile of bioengineered organisms (mainly foods).

In consideration of the heavy impact of allergic diseases on health status and quality of life of a significant part of the population, many efforts have been made to create a reliable method to predict allergenicity (i.e. the probability that a substance will cause allergy). *In vitro* and *in vivo* tests, currently used to this aim, are expensive and time consuming; additionally, obvious ethical questions arise in the case of *in vivo* tests on humans. Recent developments in the collaboration among clinicians, biochemists and bioinformaticians have created the so-called "*in silico*" approach, where calculators are used to simulate and study biologic phenomena. Although relatively "young", this method has already achieved significant results [2-5] and is considered a valid research tool that does not replace traditional techniques, but helps to better orient their use and significantly reduce the amount of resources needed.

Unfortunately, the available information about type IV immune reactions is often insufficient to create a reliable bioinformatic model. It is known that haptens can bind to several host molecules, but only few of the "new" antigens elicit a type IV response: these antigens are very often unknown, and, consequently, the definition of a common "molecular pattern" is currently almost impossible.

A different situation exists for type I reactions: allergens are, in the vast majority of cases, organic macromolecules (molecules with a high molecular weight, produced by living organisms like dust mites, plants, fungi, animals) which share a protein-type structure (proteins, glycoproteins). Proteins can be defined as "modular" structures, because they are actually chains of "molecular units" called amino acids, whose combination determines the features of each protein. This makes proteins relatively easy to study with bioinformatic techniques, at least for which concerns the amino acid sequence. Of course, function and immunogenicity of proteins do not depend only on their composition, but also on other characteristics like protein folding, that determine the three-dimensional structure. Prediction of protein structure and interaction between molecules is often an impossible task, because of the excessive number of variables involved; when it is possible, it requires high computational power and time. An easier way is amino acid sequence comparison between proteins with known three-dimensional structure and proteins whose structure is not known: although complete reliability is obviously impossible, it has been demonstrated that proteins (or parts of proteins) with a sufficiently high level of homology share similar structure and immunogenicity.

On this scientific basis, the Food and Agriculture Organization (FAO) and the World Health Organization (WHO) published in 2001 an official document entitled "Allergenicity of genetically modified foods" [6], that defined the guidelines for the *in silico* identification of possible new allergens in bioengineered foods; the rules presented are, however, also valid for non-food allergens and to search for possible allergic cross-reactivity. According to what stated in the FAO/WHO document, a potential allergen is a protein that shares with a known allergen $> 35\%$ identical amino acids in a segment of 80 consecutive amino acids (a "sliding window" of 80 amino acids), or a segment of six consecutive identical amino acids. In 2003, Stadler and Stadler tested the FAO/WHO criteria on the whole database of protein sequences available, and the results showed high sensitivity (99.7%) but low specificity (36.6%): the system classified as potential allergens 67.3% of all proteins known, 75.9% of rice proteins (although rice is not a frequent cause of allergy) and, paradoxically, even 42.9% of human proteins [7]. The insufficiently stringent criteria used, particularly for which concern the second rule, were postulated by the authors as the cause of the poor performance of the FAO/WHO method. The second rule was inserted by FAO/WHO experts to identify short allergenic proteins, but, on the other hand, the size of the segment considered (six amino acids) makes possible that amino acid identity occurs merely by chance.

Years before the FAO/WHO document, the issue of the quantification of amino acid sequence similarity between proteins had been already studied and used to evaluate the "evolutionary distance" between organisms. The most common tool in this field is BLAST (Basic Local Alignment Search Tool), a software created in 1997 by Altschul *et al.* [8]. BLAST not only finds similarities between amino acid sequences, but associates to each couple of identical or similar amino acids a similarity score, based on the frequency of occurrence of the amino acids involved. The scores are then used to compute the E value, which represents the probability that the similarity found between two proteins occurs by chance. The E value is given by the formula

$$E = Kmne^{-\lambda S} \tag{1}$$

where m and n are the numbers of amino acids in the two homologous segments, S is the minimum similarity score desired, and K and λ are statistical parameters that define the search space size and the scoring system, respectively.

After using BLAST successfully in researches on autoimmune diseases possibly caused by similarity between human and microbial proteins [3,9,10], it was natural to apply the same method to the search for new or cross-reacting allergens. In contrast to the data by Stadler and Stadler [7], results were good also in this field [11-13]. Thus, in this work it has been tested whether the use of BLAST can globally increase the specificity of the FAO/WHO criteria for the identification of allergens.

2. Materials and methods

The test was performed on a computer equipped with an Intel T2300 processor running at 1.66 GHz and 1 Gbyte of RAM memory. The BLAST software (version 2.2.18) and the non-redundant (NR) Entrez Protein database (August 2008 update) were downloaded

from the National Center for Biotechnology Information (NCBI) web site [14]. The August 2008 update of the database of allergens was downloaded from the appropriate web site owned by the International Union of Immunological Societies (IUIS) [15]. Similarly to the test performed by Stadler and Stadler [7], each allergen was compared versus the entire NR database, rice proteins, human proteins and a test database composed by three versions of each allergenic amino acid sequence (original, reversed and scrambled). In the original paper [7], the test database also contained a fourth version of each sequence, where segments of 20 amino acids each had been shuffled. The span of each segment had been chosen on the basis of the length of the shortest allergen known at the time of the study, which was made of 26 amino acids. This part of the test database was not included in the present test, because allergens shorter than 20 amino acids are now known.

All comparisons were made with BLAST, using the substitution matrix BLOSUM62, with a gap penalty of 11 for existence and 1 for extension, and without sequence masking for low complexity sequences. Each human and rice protein and each sequence in the NR database and the test database was used as a query in BLAST comparison against the allergen database. To speed up elaboration and avoid errors, we developed a software to perform the above operation automatically.

An E value < 1 is considered the minimum requirement to define a homology as significant; thus, only homologies with E <1 were selected. Next, the occurrence of the conditions required by the FAO/WHO document was researched in the selected homologies. To this aim, a software was written using Microsoft Visual Basic (version 5.0); to increase calculation speed, some parts of the program were written in Assembler, using Microsoft MASM32.

3. Results

The combination of FAO/WHO criteria and BLAST identified as potential allergens 0.51% of the $6,929,940$ proteins contained in the Entrez Protein NR database. Among the $134,043$ rice proteins, the predicted allergens are 2.09%, while the figure is 1.39% for which concerns the database of human proteins ($216,329$ in total). Table 1 compares the results of the present study with those obtained by Stadler and Stadler [7] using only FAO/WHO criteria. Also, Table 1 shows the precision of BLAST and either of the FAO rules when used individually: this underlines the important contribution of BLAST to the final results of this work. Concerning the precision of the method, the use of BLAST dramatically decreased the number of wrong predictions observed by Stadler and Stadler [7]. As shown in Fig. 1, when considering homologies with $E = 0.1$ or less, the number of "false positives" decreased to 10 (1.26% of all predictions), while all of the 785 allergens in the IUIS database were correctly identified. A slight modification of the original criteria ($> 37.5\%$ identical amino acids in a segment of 80, or 8 consecutive identical amino acids, between the protein being tested and a known allergen) achieved complete suppression of "false positives", while maintaining correct identification of allergens at $E = 0.1$ or less (see Fig. 2).

Table 1. Allergenicity predictions for all known proteins, rice proteins and human proteins: Results obtained using FAO/WHO rules only (data from Ref. [7]), BLAST only and FAO/WHO rules in conjunction with BLAST (this paper).

CRITERIA	FAO/WHO rules only	BLAST only[1]	FAO/WHO rule #1	FAO/WHO rule #2	FAO/WHO rules + BLAST[1]
Entire protein base					
Total proteins	101,602	6,929,940	6,929,940	6,929,940	6,929,940
Potential allergens	67.3%	265,863	1,428,261	4,725,526	35,501
		(3.84%)	(20.61%)	(68.19%)	(0.51%)
Rice (*Oryza sativa*)					
Total proteins	10,891	134,043	134,043	134,043	134,043
Potential allergens	75.9%	5,556	20,897	67,437	2,795
		(4.14%)	(15.59%)	(50.31%)	(2.09%)
Homo sapiens					
Total proteins	330,743	216,329	216,329	216,329	216,329
Potential allergens	42.9%	6,907	27,193	93,931	2,999
		(3.19%)	(12.57%)	(43.42%)	(1.39%)

[1] Results obtained when considering homologies with $E = 0.1$ or less.

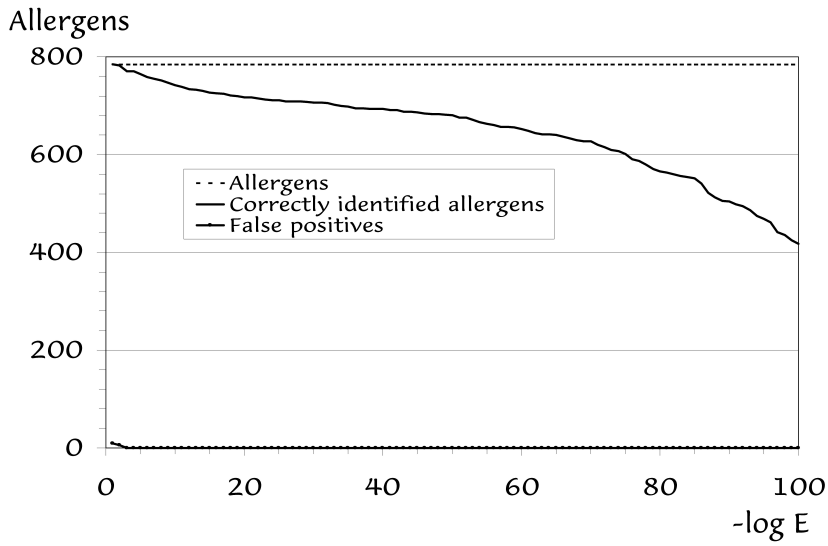

Figure 1. Allergens identified at different E values when using BLAST in combination with the original FAO/WHO criteria, i.e. either $> 35\%$ identical amino acids in a segment of 80 or six consecutive amino acids shared by a known and a candidate allergen (see Ref. 6). When $-\log E = 1$ or 2, some "false positives" (non-allergenic proteins wrongly identified as allergens) are found.

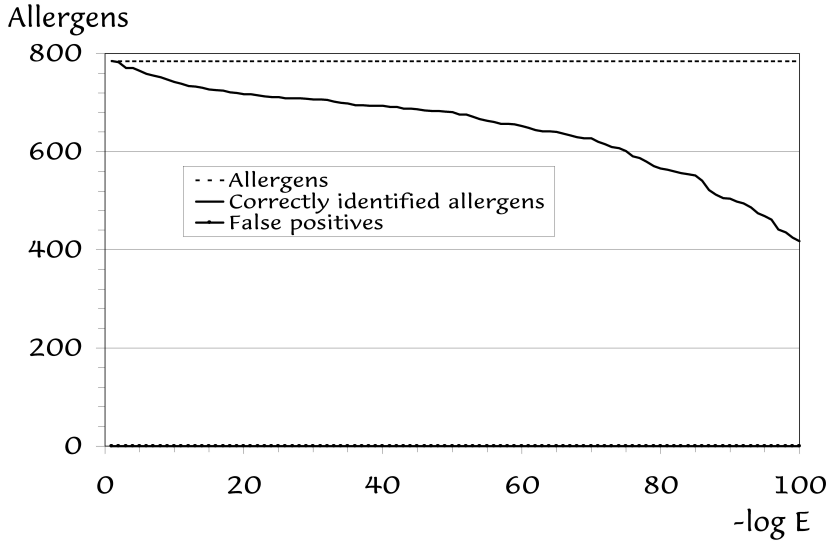

Figure 2. Allergens identified at different E values when using BLAST in combination with a modified version of FAO/WHO criteria, i.e. either $> 37.5\%$ identical amino acids in a segment of 80 or eight consecutive amino acids shared by a known and a candidate allergen. In comparison with the combined use of BLAST and original FAO/WHO criteria, the correct identification of allergens is maintained, while no "false positives" (non-allergenic proteins wrongly identified as allergens) are found.

4. Discussion

The increasing number of type I allergic reactions and their consequences - sometimes severe - on health and quality of life make research in this field an important priority. Currently used methods are efficient, but require relatively long times and are limited by economic and, in some cases, ethical reasons: thus, their large scale use for the identification of the many unknown allergens, as well as for studying the allergenicity of new proteins generated by bioengineering, is rather difficult and would require a large amount of resources. In consideration of the above difficulties, several efforts were done in the development of *in silico* research: this led to the creation of valid tools that can not replace traditional experiments, because of the extreme complexity of biologic phenomena, but can significantly narrow down the spectrum of candidate allergens, selecting only the most likely ones and, consequently, optimizing the use of the available resources.

The FAO/WHO guidelines have been maybe the first "official" attempt to introduce bioinformatics in allergologic research. Although basically valid, as demonstrated by a 99.7% rate of correct identification of allergens [7], they surprisingly did not keep into account the previous studies on the significance of protein similarity, and this made them unreliable because of an unacceptably high rate of "false positives". Conversely, successively developed methods, although based on approaches different from those outlined

in the FAO/WHO document, incorporated statistics of similarity. Among the approaches tested, noteworthy are the use of the hidden Markov model [16] and the search for "allergen motifs" [7], that both significantly improved the precision of allergen prediction.

This work shows that the use of FAO/WHO rules in conjunction with BLAST allows allergenicity prediction with a level of precision comparable, and in some cases superior, to that of the other *in silico* methods available. The improvement given by BLAST use is evident in all the tests performed: although the exact number of allergens is unknown, the prediction of 0.51% of all known proteins appears much more realistic than the 67.3% obtained with the original rules; the reliability of the proposed method is further evidenced when considering that the protein database used by Stadler and Stadler [7] included $101,602$ proteins, while the Entrez Protein NR database used in this work included $6,929,940$ sequences. Similar considerations can be made for rice allergens (2.09% of $134,043$ rice proteins are predicted in this work as potential allergens, versus a prediction of 75.9% of $10,891$ proteins in the paper by Stadler and Stadler [7]). Although rice is often recommended in hypoallergenic diets for food allergic subjects, cases of rice allergy have been reported, some allergenic proteins have been identified [17], and clinical importance of this allergy in some diseases, such as atopic dermatitis, has been demonstrated [18]. The most surprising result is probably the prediction about the allergenicity of some human proteins (1.39% of the $216,329$ known). Although apparently paradoxical, IgE reactivity against human proteins has been demonstrated in several papers [19-24], and some of the so called "autoallergens" have even been sequenced and added to the IUIS allergen database.

Little adjustments of the FAO/WHO criteria allow further improvements: the results presented in this paper show that full allergen detection, without false positives, is obtained by considering homologies with $E < 0.1$ and raising the lower limit of identical amino acids from 35% to 37.5% on an 80 amino acid long segment (i.e. from 28/80 to 30/80) for which concerns the first FAO/WHO rule, and from 6 to 8 consecutive for which concerns the second.

Notwithstanding the above results, we think that the method proposed here is a significant improvement, but cannot be considered the definitive solution in the field of *in silico* allergen identification. The predicted number of allergens is $35,501$, while the most updated allergen databases (e.g. the Food Allergy Research and Resource Program (FARRP) [25]) report about $1500 - 1800$ known allergens (including some experimentally confirmed allergens that do not fulfill the requirements for inclusion in the IUIS database). While some data suggest that the majority of allergens is yet to be characterized (for example, statistics of the Allergome database [26], report that 994 out of 1748 known allergen sources still have unidentified allergenic molecules), it is reasonable to suppose that the above prediction overestimates the actual number of allergens, by an amount that can be precisely defined only experimentally. Another possible issue, shared by all homology-based prediction systems, is that while many allergens are similar among themselves, others with an "unusual" structure are sometimes discovered, and the performance of existing softwares in such cases is unpredictable. It is then necessary to test the available algorithms on larger allergen databases and to establish a close collaboration between clinical and bioinformatic researchers, to achieve a better "tuning" of rules and parameters.

References

[1] P. G. H. Gell and R. R. A. Coombs, *Clinical aspects of immunology*, 1st ed. (Blackwell, Oxford, 1963).

[2] A. Persidis, "Autoimmune disease drug discovery", *Nat. Biotechnol.* **17**, 1038 (1999).

[3] S. Benvenga, F. Guarneri, M. Vaccaro, L. Santarpia, and F. Trimarchi, "Homologies between proteins of *Borrelia burgdorferi* and thyroid autoantigens", *Thyroid* **14**, 964 (2004).

[4] W. F. Nieuwenhuizen, R. H. H. Pieters, L. M. J. Knippels, M. C. J. Jansen, and S. J. Koppelman, "Is *Candida albicans* a trigger in the onset of coeliac disease? ", *Lancet* **361**, 2152 (2003).

[5] C. Lunardi, C. Bason, M. Leandu, R. Navone, M. Lestani, E. Millo, U. Benatti, M. Cilli, R. Beri, R. Corrocher, and A. Puccetti, "Autoantibodies to the inner ear and endothelial antigens in Cogans syndrome", *Lancet* **360**, 915 (2002).

[6] FAO/WHO, *Allergenicity of genetically modified foods* (FAO/WHO, Rome, 1999).

[7] M. B. Stadler and B. M. Stadler, "Allergenicity prediction by protein sequence", *FASEB J.* **17**, 1141 (2003).

[8] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.* **25**, 3389 (1997).

[9] S. Benvenga, L. Santarpia, F. Trimarchi, and F. Guarneri, "Human thyroid autoantigens and proteins of *Yersinia* and *Borrelia* share amino acid sequence homology that includes binding motifs to HLA-DR molecules and T-cell receptor", *Thyroid* **16**, 225 (2006).

[10] F. Guarneri, C. Guarneri, and S. Benvenga, "*Helicobacter pylori* and autoimmune pancreatitis: role of carbonic anhydrase via molecular mimicry? ", *J. Cell. Mol. Med.* **9**, 741 (2005).

[11] F. Guarneri, C. Guarneri, and S. Benvenga, "Cross-reactivity of *Anisakis simplex*: possible role of Ani s 2 and Ani s 3", *Int. J. Dermatol.* **46**, 146 (2007).

[12] F. Guarneri, C. Guarneri, B. Guarneri, and S. Benvenga, "In silico identification of potential new latex allergens", *Clin. Exp. Allergy* **36**, 916 (2006).

[13] F. Guarneri, C. Guarneri, and S. Benvenga, "Identification of potentially cross-reactive peanut-lupine proteins by computer-assisted search for amino acid sequence homology", *Int. Arch. Allergy Immunol.* **138**, 273 (2005).

[14] Available at http://www.ncbi.nlm.nih.gov/; accessed 12 September 2010.

[15] Available at http://www.allergen.org/; accessed 12 September 2010.

[16] K. B. Li, P. Issac, and A. Krishnan, "Predicting allergenic proteins using wavelet transform", *Bioinformatics* **20**, 2572 (2004).

[17] Y. Usui, M. Nakase, H. Hotta, A. Urisu, N. Aoki, K. Kitajima, and T. Matsuda, "A 33-kDa allergen from rice (*Oryza sativa L. Japonica*). cDNA cloning, expression, and identification as a novel glyoxalase I", *J. Biol. Chem.* **276**, 11376 (2001).

[18] Z. Ikezawa, K. Miyakawa, H. Komatsu, C. Suga, J. Miyakawa, A. Sugiyama, T. Sasaki, H. Nakajima, Y. Hirai, and Y. Suzuki, "A probable involvement of rice allergy in severe type of atopic dermatitis in Japan", *Acta Derm. Venereol. Suppl. (Stockh.)* **176**, 103 (1992).

[19] R. Valenta, S. Natter, S. Seiberler, S. Wichlas, D. Maurer, M. Hess, M. Pavelka, M. Grote, F. Ferreira, Z. Szepfalusi, P. Valent, and G. Stingl, "Molecular characterization of an autoallergen, Hom s 1, identified by serum IgE from atopic dermatitis patients", *J. Invest. Dermatol.* **111**, 1178 (1998).

[20] R. Valenta, S. Natter, S. Seiberler, M. Roschanak, N. Mothes, V. Mahler, and P. Eibensteiner, "Autoallergy: a pathogenetic factor in atopic dermatitis? ", *Curr. Probl. Dermatol.* **28**, 45 (1999).

[21] S. Seiberler, S. Natter, P. Hufnagl, B. R. Binder, and R. Valenta, "Characterization of IgE-reactive autoantigens in atopic dermatitis. 2. A pilot study on IgE versus IgG subclass response and seasonal variation of IgE autoreactivity", *Int. Arch. Allergy Immunol.* **120**, 117 (1999).

[22] R. Valenta, S. Seiberler, S. Natter, V. Mahler, R. Mossabeb, J. Ring, and G. Stingl, "Autoallergy: a pathogenetic factor in atopic dermatitis? ", *J. Allergy Clin. Immunol.* **105**, 432 (2000).

[23] Y. Muro, "Autoantibodies in atopic dermatitis", *J. Dermatol. Sci.* **25**, 171 (2001).

[24] P. Schmid-Grendelmeier, S. Flückiger, R. Disch, A. Trautmann, B. Wüthrich, K. Blaser, A. Scheynius, and R. Crameri, "IgE-mediated and T cell-mediated autoimmunity against manganese superoxide dismutase in atopic dermatitis", *J. Allergy Clin. Immunol.* **115**, 1068 (2005).

[25] Available at http://www.allergenonline.org/; accessed 12 September 2010.

[26] Available at http://www.allergome.org/; accessed 12 September 2010.

[a] Università degli Studi di Messina
Dipartimento di Medicina Sociale del Territorio, Sezione di Dermatologia,
A.O.U. Policlinico "G. Martino", Pad. H, Piano 4,
Via Consolare Valeria - Gazzi -
98125 Messina, Italy

E-mail: fguarneri@unime.it

Presented 25 November 2009; published online 5 October 2010