doi: 10.13241/j.cnki.pmb.2014.35.053

与疾病相关的非同义单核苷酸多态性预测的研究进展 *

喻海燕[1] 赵 健[1] 张珊珊[1] 韩 平[2] 宋晓峰[1△]

（1 南京航空航天大学生物医学工程系 江苏 南京 210016；2 南京医科大学第一附属医院 江苏 南京 210029）

摘要：单核苷酸多态性（single nucleotide polymorphism, SNPs），即在基因组水平上由单个核苷酸的变异而引起的 DNA 序列多态性变化，具体是指在 DNA 序列中的单个碱基的变异，其是人类基因组变异种最常见的一种。SNP 研究最主要的目的就是对人类表型变异遗传学的理解，尤其是关于人类遗传疾病的研究。而非同义单核苷酸多态性（nsSNPs）是 SNPs 中的一种，主要是指处于编码区会引起翻译后对应氨基酸序列变化的单核苷酸突变。因为 nsSNPs 可能会对蛋白质的功能造成影响，被认为是造成人类遗传病的主要原因。因此将与疾病相关的 nsSNPs 从中性的 nsSNPs 中区分出来是很重要的。本文根据国内外与疾病相关 nsSNPs 预测的研究，分析了预测中所涉及到的特征属性，总结了对这些特征进行优化的特征选择方法，并概述了在预测过程中使用的各种分类器。

关键词：非同义单核苷酸多态性；nsSNPs 预测；特征选择；分类器

中图分类号：Q75；TP181 **文献标识码**：A **文章编号**：1673-6273（2014）35-6996-05

Progress in Prediction of Disease-Associated Non-Synonymous Single Nucleotide Polymorphisms*

YU Hai-yan[1], ZHAO Jian[1], ZHANG Shan-shan[1], HAN Ping[2], SONG Xiao-feng[1△]

(1 Department of Biomedical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, 210016, China; 2 Affiliated Hospital of Nanjing Medical University, Nanjing, Jiangsu, 210029, China)

ABSTRACT: SNPs stands for single nucleotide polymorphism, which are single-nucleotide substitutions of one base for another in the DNA sequence which represents the most frequent type of human population DNA variation. One of the most important motivations to do SNPs research is to help understand the genetics of the human phenotype variation and especially the genetic base of complex diseases. Non-synonymous single nucleotide polymorphisms (nsSNPs) are coding mutation that introduces amino acid changes in their corresponding proteins. nsSNPs are believed to be the main cause of human genetic disease because they can affect protein function. Therefore, it is important to distinguish the disease-related nsSNPs from those are neutral. Based on domestic and international research about the prediction of disease-associated nsSNPs, the essay analyzed the characteristic attributes involved in prediction and summarized the attribute selection methods that can extract the informative features improving the performance of classifier, meanwhile, an overview of the various classifiers used in the prediction process.

Key words: Non-synonymous single nucleotide polymorphisms; Prediction of disease-associated nsSNPs; Attribute selection methods; Classifiers

Chinese Library Classification: Q75; TP181 Document code: A

Article ID: 1673-6273(2014)35-6996-05

前言

随着人类基因组研究计划（HGP）DNA 序列测定工作的迅速发展和二代测序的大量使用，发现了很多没有预期得到的基因变异，研究人类基因组变异也相继变得越来越受到人们的重视。基因组变异中最常见的一种形式就是单核苷酸多态性（single nucleotide polymorphism, SNPs）。单核苷酸多态性，即在基因组水平上由单个核苷酸的变异而引起的 DNA 序列多态性变化，具体是指在 DNA 序列中的单个碱基的变异。其分为两种形式：一种是位于编码区引起翻译后的氨基酸序列变化发生改变从而导致蛋白质功能发生改变的 cSNPs，即错义突变（Non-synonymous single nucleotide polymorphisms，nsSNPs）；另一种即不会造成氨基酸序列发生改变的突变即同义突变。错义突变是造成人类遗传病的主要原因。SNPs 其最早的功能是作为一种基因组作图的遗传标记，随着研究的进一步深入，发现 SNPs 还有利于复杂性疾病相关基因的研究和药学基因组学的

* 基金项目：国家自然科学基金项目（61171191）；江苏省自然科学基金项目（BK2010500）

作者简介：喻海燕（1988-），女，硕士研究生，研究方向：生物信息学，E-mail：yanhaiyuzong@sina.com

△通讯作者：宋晓峰，Email：xfsong@nuaa.edu.cn

（收稿日期：2014-05-27 接受日期：2014-06-21）

发展，因此预测与疾病相关的 nsSNPs 也受到了人们的广泛关注。本文就国内外预测与疾病相关的 nsSNPs 的研究进展做一综述，并对其提出展望。

1 概述

在 dbSNP 数据库中 [1] 收录着数百万的 SNPs，在数据库 OMIM(Online Mendelian Inheritance in Man)[2]和 HGMD(Human gene mutation database)[3]中收录了一些于疾病相关的突变记录并且显示大部分引起疾病的突变都是错义突变。这些突变可能会影响蛋白质的功能，例如改变转录，RNA 转录后修饰，蛋白质表达，多肽链的折叠，折叠状态的稳定性，翻译后修饰和催化等方面。这些改变造成功能的改变，从而造成与该蛋白相关的疾病。例如 C-subunit(c AMP-dependent protein kinase catalytic subunit) 的激活位点处有一个由几个非保守的氨基酸残基（Glu127, Glu170, Glu203, Glu230, and Asp241）组成的一个集对于起底物识别和连接是至关重要的，研究发现 Glu230 突变成Gln 将减少该酶对于底物肽的亲活性从而破坏其自身的功能 [4]；在因子 VII 中的突变 Arg304Gln 导致血浆 FVII 的活动变得无法预测，表明这个残基与其周围的结构可能对这个蛋白质的功能起着重要的作用 [5]。因此从大量的 SNPs 中将与疾病有关的 nsSNPs 使用生物信息学方法识别出来成为了研究热点。目前国内外的很多学者在预测与疾病相关的 SNPs 方面已经取得一定的成果了，这里我们将列举其中的部分成果[6-9]，如表 1 所示。

表 1 几种预测与疾病相关的 nsSNPs 研究成果

Table 1 Studies about prediction of disease-associated nsSNPs

Method	Interface	Performance	Feather selected	Algorithm
TopoSNP[6] http://gila.bioengr.uic.edu/snp/toposnp	Input: Protein id or protein sequence Output: Can view position of mutation. Location of substitution on protein (surface, internal, or pocket) and conservation reported separately Results are stored so an input protein sequence not in the database will not be processed	FN error: 12% FP error: NA Coverage: NA	Base on protein structure	Classifies substitution as buried, on the surface, or in a pocket of the protein's structure.
SIFT[7] http://blocks.fhcrc.org/sift/SIFT.html	Input: Protein sequence and AAS, Protein sequence alignment and AAS, dbSNP id, or protein id Output: Score ranges from 0 to 1, where 0 is damaging and 1 is neutral	FN error: 31% FP error:20% Coverage: 60%	Base on protein sequence homology	position-specific scoring matrices with Dirichlet priors
PolyPhen[8] http://www.bork.embl-heidelberg.de/PolyPhen	Input: Protein sequence and AAS, dbSNP id, HGVbASE id, or protein id Output: Score ranges from 0 to a positive number, where 0 is neutral, and a high positive number is damaging	FN error: 31% FP error:9% Coverage: 81%	Base on protein structure and function and multiple alignment	Uses sequence conservation, structure to model position of amino acid substitution, and SWISS-PROT annotation
SNPs3D[9] http://www.snps3d.org/	Input: dbSNP id, protein id, literature search, or gene ontology Output: Scores from structure-based SVM and sequence-based SVM reported separately. Score <0is amaging. Mutation on protein structure can be visualized	Structure-based FN error: 26% FP error: 15% Coverage: 14% Sequence-based FN error: 20% FP error: 10% Coverage: 71%	Base on protein structure or protein sequence homology	Structure-based support vector machine uses 15 structural factor Sequence-conservation support vector machine uses five features that capture sequence conservation

从上述研究中我们可以看出，各位研究学者所采用的特征属性的来源有不同也有重叠的部分，建立预测模型的方法有一定的差别，得到的预测结果也比较理想。在这里对与疾病相关的单核苷酸多态性（SNPs）预测的研究进展基于所选取的特征值方面和建模方面的研究做一个概述。

2 预测与疾病相关 nsSNPs 的相关特征分析

2.1 序列方面特征

从序列方面进行的分析主要有氨基酸理化性质、突变计分矩阵、多重序列比对和序列派生信息，下面将对此进行详细介

绍。

通过分析突变前后的氨基酸的理化性质的变化将对研究与疾病相关的 nsSNPs 有重要的作用，而每一种氨基酸的理化特性都能用 20 个数值组成的数据集来表示，这 20 个数值即称为氨基酸指数。AAIndex[10]是 Shuichi Kawashima et al 开发出来的一些通过大量实验和理论研究确定的氨基酸的理化特性构成的数据库，其包含氨基酸残基或者是氨基酸对的各种生化和物化特性。在预测与疾病相关的 nsSNPs 研究中，存在理化特性冗余的问题，因为有些理化特性是与研究不相关的，将可能影响预测的准确性，所以一般都会进行特征选择[11]。

蛋白质序列计分矩阵即记录序列比对时两个相对应的残基的相似度，当矩阵定义好后，比对程式就可以利用这个矩阵，尽量将相似的残基排在一起，以达到最好的比对。现在用的比较多的计分矩阵有两种 PAM[12]和 BLOSUM[13]。PAM 是基于进化的点突变模型，如果两种氨基酸替换频繁，说明了这种突变可以被接受，那么这对氨基酸替换得分就高。而 BLOSUM 是采用同源与非同源的可能性的比率的对数来打分。在 Carles Ferrer-Costa et al.[14] 的研究中对这两种方法都加以采用并且表明，与疾病相关的 nsSNPs 更倾向分布于打分值为负值，而中性的 nsSNPs 则倾向于分布于打分值大于 0. 因此突变打分矩阵可能会对于区分中性和与疾病相关的 nsSNPs 起到一定的作用。

通过多重序列比对，如果序列中某个位置是高度保守的，那么它有可能就是功能性位点，该位点发生突变，将可能会对蛋白质的功能造成影响。表 1 中 SIFT[7]工具就是完全使用序列同源的方法来预测氨基酸替换是否会影响蛋白质的功能。Pauline C.Ng et al.[7] 在其研究中指出替换打分矩阵可能会低估了处于关键位置氨基酸替换的严重性。而在某些位置上，如果该位置并不参与蛋白质的功能和构成，那么蛋白质中该氨基酸的替换是可以被允许的，为中性的替换。因此使用基于同源序列的方法将有效的预测一个特殊位置的氨基酸替换是否会导致表型的变化，提高预测的正确性。研究发现中性的突变发生频率较高，其一般是出现在低保守性位置，而与疾病相关的 nsSNPs 的突变发生频率虽然较低，其主要是出现在高保守性的位点上。但是当采用多重序列比对方法时预测的正确性依赖于足够数量的同源序列。Saunders 和 Baker[15]在其研究中证明当同源序列数目少于 5-10 条时预测的准确率将显著下降。

当一个 nsSNPs 的位置是处于一个功能性域或是功能性位点周围，那么这个 nsSNPs 就很有可能是与疾病有关系的。所以在研究中，许多学者都加入了功能性位点这一特征 [16-19]。在 Swiss-prot[20]数据库中记录了人类蛋白质已实验验证的所有功能性位点，包括激活位点（ACT_SITE）、连接位点（BINDING）、金属离子结合位点（METAL）、翻译后修饰位点（MOD_RES）、二硫键（DISULFID）和跨膜区（TRANSMEM）。

2.2 结构方面特征

从序列方面进行预测的缺点是它不能直接的洞察 nsSNPs 对蛋白质造成的影响，因此在很多学者采用结构方面的特征或者是将结构与序列结合起来进行分析。在 Sunyaev S et al.[21]的研究中发现，与疾病相关 nsSNPs 更倾向于分布在溶剂可及性 <5%或者 β 折叠。

一般 75%的氨基酸残基可以被替换而不改变蛋白质的结构，然而有时改变单个关键的氨基酸残基则可能导致蛋白质的结构破坏。蛋白质的二级结构对整个蛋白折叠的稳定性起着非常重要的作用，从而进一步影响其功能。二级构象信息可以从 HSSP 文件中提取得到，细分为：处于孤立的 β-bridge 中的残基，5/10 helix，3/10 helix，4/10 helix, bend, β-sheet。

蛋白质溶剂可及性是描述蛋白质疏水性的重要手段，蛋白质分子中残基的疏水性是影响蛋白质折叠的重要物理作用，并对蛋白质的空间构象以及构象的柔性有重要的影响。残基溶剂可及性有两种，一种是该氨基酸所有原子的溶剂可及性之和，另一种是 C_β 密度的计算。在 Zhi-Qiang Ye et al.[16] 的研究中分析了不同定义的溶剂可及性，突变前和突变后的残基的所有原子，所有侧链，主链，非极性侧链和所有极性侧链都分别计算了其绝对的和相对的溶剂可及性。 类似的在 Nathan O.Stitziel et al.[22] 的研究中，则是通过 nsSNPs 在蛋白质结构中的定位来进行分析，并且表明与疾病相关的 nsSNPs 更倾向于分布在表面。

研究发现，与疾病相关的 nsSNPs 例子中 80%以上会造成蛋白质结构的不稳定。影响蛋白质稳定性的因子有：电荷 - 电荷、电荷 - 极性或者极性 - 极性能量的减少，或者是静电斥力的引进；电荷或极性集的掩埋，或者埋藏在折叠区的非极性区域的减少；骨架的变化即拉紧或重叠；还有就是影响范德华力和二硫键的丢失。在预测时候这些特征都经验规则的作用。

2.3 靶蛋白的交互网络

传统的预测都是基于结构或者序列特征的，因为他们只是关注于蛋白质本身的变化。仅仅一个 nsSNPs 靶蛋白的变化就能决定或者造成一种病理生理表型这并不具有足够的说服力，因此在 Tao Huang et al.[23]的研究中，他们加入了蛋白质交互网络这一特征。在一个网络中，一些结点位于很重要的位置，其他的必须依赖这些结点来交换信息。每一个结点的网络可以使用 Freeman[25]的中间中心度方法来研究。中间中心度衡量的是一个行动者在他相连的行动者之间所起到的影响和控制作用。在一个图 G=(V,E)中，结点的中间度被定义为：

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\delta_{st}(v)}{\delta_{st}}$$

s 和 t 是网络中的其他结点，δ_{st} 为结点 s 和结点 t 之间最短路径的数目，$\delta_{st}(v)$ 是经过 v 的路径的数目。如果该行动者位于 n 对行动者之间的最短路径上，n 越大，说明该行动者的中间中心度越大，也就是他的媒介能力越强。

3 特征选择方法

从上面的研究中我们可以看出，对于特征的提取可以从很多方面进行，我们不明确是不是所有的特征都能为我们的预测提供依据。对选取得到的特征进行去冗余可以提高分类器预测的准确性和性能，从而产生高效率低成本的分类器，因此特征选择是很重要的。特征选择的方法有很多种，例如 F 检验、T 检验、最大冗余最小相关算法（mRMR）、增量特征选择方法（IFS）等。

F 检验和 T 检验都是常用的能有效区分两个样本集的测量方法,主要应用统计学的相关原理。不同的是 F 检验用于区分两个样本集的测量方法,F 值越高则说明该特征区分正样本和负样本的能力越强。而 T 检验是检验一个总体中的小样本平均数是否等于总体平均值的检验方法,得到的 P 值是将观测结果认为具有总体代表性的犯错概率,P 值越小则说明该特征的统计显著性越强,对预测性能的提高越有利。

最大冗余最小相关算法（Maximum Relevance, Minimum Redundancy , mRMR)最开始是由 Peng et al.[26] 提出,mRMR 软件包可以从网上下载得到。其基本思想是根据目标变量的相关性和特征间的冗余性对特征进行排序,若一个特征与目标变量具有最大相关且在特征间有最小冗余,则该特征的区分能力最强。

我们在经过特征选择后,得到了一系列具有区分能力的特征列表,但是不知道列表中应该选择多少个特征,在这时就要用稿增量特征选择法。增量特征选择法（Incremental Feature Selection, IFS)[27,28], 即首先用于分类器无关的评价函数选出候选特征集合,然后将分类算法作用于候选特征集合,利用分类精度作为评价标准去选择特征子集,在遇到概念漂移时重新选择特征子集。在 Tao Huang 等人[23]的研究中使用 mRMR 方法先对特征进行排序选取出所有具有区分能力的特征,然后使用增量特征选择的方法(IFS)来决定最终多少个特征最终被选择。

4 与疾病相关的 nsSNPs 的分类方法

4.1 基于经验规则的分类方法

Wang 和 Moult (2001) [24]年的研究表明大多数有害的 nsSNPs 都是间接的通过影响蛋白质结构稳定性来对蛋白质的功能造成影响,例如蛋白质疏水核心的中断等。随后的学者们在此的基础上提供了一系列的经验规则来预测有害的 nsSNPs [15,17,29,30]。Christopher T.Saunders 和 David Baker[30]的研究中使用的是基于蛋白质结构的经验规则,包括:(1)疏水核心的中断:即突变位点的溶剂可及性低于 25%并且突变的两个残基之间的可接触表面倾向值的差异大于 0.75;(2)掩埋电荷的变化:即突变位点的溶剂可及性低于 25%并且突变造成了静电电荷的改变;(3)溶解度的变化:突变位点的溶剂可及性大于 50%并且突变的两个残基之间可接触表面倾向值的差异大于 2.0;(4)α-helix 中的脯氨酸:任何突变成脯氨酸的突变位点通过 DSSP 预测是位于一个 α-helix 中。在其研究中预测达到了比较好的的效果,交叉验证错误率为 20.5%。

4.2 基于突变位点进化信息的分类方法

Pauline C.Ng et al.(2001 年）的基于序列保守性 SIFT[7] (Sorting Tolerant from Intolerant)方法主要是通过 BLAST 搜索程序和多序列比对技术来评价和预测点突变对蛋白质功能的影响。其首先通过 BLAST 程序搜索和目标序列相似的序列,然后从搜索得到的相似序列中通过序列比对的方法获得在编码和功能上与目标序列最接近的序列,组成序列组;最后使用多重序列比对方法分析序列组,计算每一个位点保守性,获得每一个氨基酸在特定位点上对蛋白质结构产生影响的概率,并根据此概率确定突变对蛋白质功能的影响程度,以及通过突变出现的频率和发生突变位点的性质来评价突变对蛋白功能的影响。标准方差小于 0.05 的位点突变被认为是有害的,大于或等于 0.05 的位点突变是可以允许的变异。

类似的还有 Robert J. Clifford et al 基于蛋白质家族信息的 Pfam-based LogR.E-value 法[31,32],该方法主要关注于在进化过程中保守的 motif 中的氨基酸替换。使用了 HMMER 软件包[33]和 Pfam 属性[34,35],其对每个氨基酸替换对蛋白质序列中 motif 的影响值。基于 pfam 蛋白质模式模型提出假设如果氨基酸的替换减少了蛋白质序列到域模型的匹配度,那么它很可能对蛋白质的活动或者是稳定性造成了影响,通过 HMMER2 程序包来量化每个替换影响匹配度的, 产生的统计值中的 E-value 代表预测中随机生成且比查询序列匹配度高的序列的数目。根据标准的和突变后的蛋白质序列计算出一个度量标准 LogR.E-value。LogR.E-value 值大于表明氨基酸的替换减少了序列到 Pfam 模式模型的匹配度。该研究表明由氨基酸替换引起 LogR.E-value 值显著变化可以用于很好的预测与疾病相关的 nsSNPs。

4.3 机器学习方法

目前被广泛使用还有一种分类方法即机器学习分类器,为每个突变定义一系列描述属性后构建一个由大量参数刻画的模型,由器自动处理数据,使信息提取过程尽可能地实现自动化。机器学习方法近似最优,也更容易执行严格的交叉验证。在预测与疾病相关的 nsSNPs 研究中常用的机器方法有决策树和支持向量机(SVM),随机森林(RF)等。

支持向量机(Support Vector Machine,SVM)是 Vapnik[36]提出的以结构风险最小化原理为基础的分类方法。该方法最初来自于对二值分类问题的处理,其机理是将数据映射到一个高维的特征空间中,寻找一个具有最大间隔的分割超平面使得学习器得到全面最优化。Chih-Jen Lin et al. 开发了一个易于操作、快速有效的 SVM 软件包 LIBSVM[37],可以供用户解决各种分类问题。其方法通常是将训练的数据分为两部分,一部分用来实现 SVM 的训练(training set),一部分用来对训练好的 SVM 进行测试(test set),测试的准确度可以很好的反映 SVM 对未知数据的预测性能。在 Zhi-Qiang[16]等人的研究中就使用了 SVM 的方法作为分类器达到了很好的预测效果, 其准确率达到了 82.61%。

决策树分类算法是一种逼近目标功能离散值的方法。它为一种典型的分类方法,首先对数据进行处理,利用归纳算法生成可读的规则和决策树,然后使用决策对数据进行分析,其本质是通过一系列规则对数据进行分类的过程。每个实例(在这里是指一个 nsSNPs）根据其属性的值（例如涉及的残基的类型、序列保守值,结构特征等)从根节点开始排序直到它到达一个分类叶子结点(有效或是无效),就完成了它的预测。当前最有影响力的决策树算法是 Quinlan 于 1986 年提出的 ID3 [38]和 1993 年提出的 C4.5[39],其中 C4.5 是 ID3 的改进算法,不仅可以处理离散型描述属性,还能处理连续性描述属性。决策树软件中为每条规则提供了一个源于训练数据的估计的精确度,这些

估计的精确度用于分配预测的置信水平。

随机森林分类器(RF)是一组由多个决策树构成,每个决策树都依赖独立采样的向量值,在随机森林中所有的决策树都具有相同的分布,它利用了机器学习算法 bagging[40]和 随机特征选择的优势。在 bagging 算法中,每棵树都是在训练数据的引导样本上进行训练,预测是由占优势的决策树来决定。随机森林可以处理大量的输入变量、学习快速,对于不平衡的分类数据集其还可以平衡误差。在 Lei bao 和 Yan Cui[41]的研究中使用了 RF 和 SVM 两种分类器分别进行预测,研究表明随机森林的结果比 SVM 的预测具有更好的性能,可能原因是 RF 在处理相关预测因子的能力方面比 SVM 更好。

5 问题与展望

SNP 影响蛋白质功能的研究已成为一个十分活跃的研究领域了,本文通过概述表明与疾病相关的 nsSNPs 与蛋白质序列中氨基酸的理化性质、进化信息,蛋白质家族信息、结构特性、功能特征等都存在着一定的关系。研究有害的 nsSNPs 的关联特征,可以帮助我们有效的将与疾病相关的 nsSNPs 从中性的 nsSNPs 中区分出来。虽然关于有害的 nsSNPs 的研究已经取得一定的成果,通过不断的改进方法预测的准确率已经获得逐渐的提高。但是还是存在着一定的问题例如有的预测是基于蛋白质的结构的,而现有的数据库中通过实验验证的蛋白质的结构仅仅只是占了一部分,所以很多结构模型都是通过建模软件得到的,这样降低了可信度。但是随着实验水平的提高,通过实验验证的蛋白质的结构也将越来越多,蛋白质结构数据库中的数据也将越来越完善,这将为我们的研究提供强大的数据支持。

参 考 文 献(References)

[1] Smigielski E M, Sirotkin K, Ward M, et al. dbSNP: a database of single nucleotide polymorphisms [J]. Nucleic Acids Research, 2000, 28(1): 352-355

[2] Hamosh A, Scott A F, Amberger J S, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders [J]. Nucleic acids research,2005,33 (suppl 1): D514-D517

[3] Stenson P D, Ball E V, Mort M, et al. Human gene mutation database (HGMD®): 2003 update[J]. Human mutation, 2003,21(6):577-581

[4] Ung M U, Lu B, McCammon J A. E230Q mutation of the catalytic subunit of cAMP-dependent protein kinase affects local structure and the binding of peptide inhibitor[J]. Biopolymers,2006,81(6): 428-439

[5] Ferrer-Costa C, Orozco M, de la Cruz X. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties [J]. Journal of molecular biology, 2002,315(4):771-786

[6] Stitziel N O, Binkowski T A, Tseng Y Y, et al. topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association [J]. Nucleic acids research,2004,32(suppl 1):D520-D522

[7] Ng P C, Henikoff S. SIFT: Predicting amino acid changes that affect protein function[J]. Nucleic acids research,2003,31(13):3812-3814

[8] Adzhubei I A, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations [J]. Nature methods,2010,7 (4):248-249

[9] Yue P, Melamud E, Moult J. SNPs3D: candidate gene and SNP selection for association studies [J]. BMC bioinformatics,2006,7(1): 166

[10] Kawashima S, Kanehisa M. AAindex: amino acid index database[J]. Nucleic acids research,2000,28(1):374

[11] Atchley W R, Zhao J, Fernandes A D, et al. Solving the protein sequence metric problem[J]. Proceedings of the National Academy of Sciences of the United States of America,2005,102(18):6395-6400

[12] Dayhoff M O, Schwartz R M. A model of evolutionary change in proteins[C]//In Atlas of protein sequence and structure. 1978

[13] Henikoff S, Henikoff J G. Amino acid substitution matrices from protein blocks [J]. Proceedings of the National Academy of Sciences, 1992,89(22):10915-10919

[14] Ferrer-Costa C, Orozco M, de la Cruz X. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties [J]. Journal of molecular biology, 2002,315(4):771-786

[15] Saunders C T, Baker D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction [J]. Journal of molecular biology,2002,322(4):891-901

[16] Ye Z Q, Zhao S Q, Gao G, et al. Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP)[J]. Bioinformatics, 2007,23(12):1444-1450

[17] Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey[J]. Nucleic acids research, 2002, 30(17):3894-3900

[18] Hu J, Yan C. Identification of deleterious non-synonymous single nucleotide polymorphisms using sequence-derived information [J]. BMC bioinformatics,2008,9(1):297

[19] Dobson R, Munroe P, Caulfield M, et al. Predicting deleterious nsSNPs: an analysis of sequence and structural attributes [J]. BMC bioinformatics,2006,7(1):217

[20] Boeckmann B, Bairoch A, Apweiler R, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003 [J]. Nucleic acids research, 2003,31(1):365-370

[21] Sunyaev S, Ramensky V, Bork P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms[J]. Trends in genetics: TIG,2000,16(5):198

[22] Stitziel N O, Tseng Y Y, Pervouchine D, et al. Structural location of disease-associated single-nucleotide polymorphisms [J]. Journal of molecular biology, 2003,327(5):1021-1030

[23] Huang T, Wang P, Ye Z Q, et al. Prediction of deleterious non-synonymous SNPs based on protein interaction network and hybrid properties[J].PLoS One,2010,5(7): e11900

[24] Wang Z, Moult J. SNPs, protein structure, and disease [J]. Human mutation,2001,17(4):263-270

[25] Freeman L C. Centrality in social networks conceptual clarification [J]. Social networks,1979,1(3):215-239

(下转第 6979 页)