

后基因组时代生物信息学的新进展*

王正华, 王勇献

(国防科技大学计算机学院, 湖南长沙 410073)

摘要: 生物信息学是近年来综合生命科学与信息科学而迅速发展起来的一门新兴交叉学科。对后基因组时代生物信息学的几个热点研究方向进行了概述, 介绍了当前一组有影响的代表性工作, 初步分析了各种研究的技术特点与适用范围, 并对今后的研究作了展望, 提出了亟待解决的一些关键问题。

关键词: 生物信息学; 基因组信息学; 蛋白质组学

中图分类号: Q 811; TP391 **文献标识码:** A

Progress of Bioinformatics in Post-genome Era

WANG Zheng-hua, WANG Yong-xian

(College of Computer, National Univ. of Defense Technology, Changsha 410073, China)

Abstract With the rapid development of biotechnology and informatics, bioinformatics emerges. An introduction and overview of the current state of the field of bioinformatics are proposed and some representative studies are also introduced and analyzed. The characteristics of current and future research are summarized and a few of questions to be resolved urgently are proposed finally.

Key words: bioinformatics; genome informatics; proteomics

生物信息学 (bioinformatics) 是在生命科学、计算机科学和数学的基础上逐步发展而形成的一门新兴交叉学科, 是为理解各种数据的生物学意义, 运用数学与计算机科学手段进行生物信息的收集、加工、存储、传播、分析与解析的科学^[1~3]。由于历史原因, 有的研究者也使用计算生物学 (computational biology) 或计算分子生物学 (computational molecular biology) 等不同的术语。生物信息学的产生最早可以追溯到 1956 年在美国田纳西州的 Gatlinburg 召开的首次“生物学中的信息理论讨论会”^[4], 但直到上个世纪八九十年代才得到快速发展。20 世纪末, 随着各生物物种基因组研究的相继开展, 序列测定技术日益向工业化发展, 基因重组、基因芯片和多维核磁共振等各项新技术也得到广泛应用, 生物学实验数据呈爆炸趋势增长; 另一方面, 计算机信息技术和国际互联网的发展也使得对大规模数据的收集、存储、处理和解析成为可能, 生物信息学因此得以迅速发展, 并日益显示出强大的生命力。

生物信息学自产生以来, 大致经历了前基因组时代、基因组时代和后基因组时代三个发展阶段。前基因组时代的标志性工作包括生物数据库的建立、检索工具的开发以及 DNA 和蛋白质序列分析等; 基因组时代的标志性工作包括基因识别与发现、网络数据库系统的建立和交互界面工具的开发等; 后基因组时代的标志则是大规模基因组分析、蛋白质组分析以及各种数据的比较与整合。这三个阶段虽无明显的界限, 但能反映出整个研究重心的转移变化情况。在后基因组时代, 生物信息学的研究内容主要可分为两个重要组成部分: 基因组信息学和蛋白质组信息学。

1 基因组信息学的研究进展

基因组信息学 (genome informatics) 的主要任务是根据各种生物基因组和测序的资料与数据, 应用数学与统计学、计算机科学、遗传学与分子生物学等多种学科中的研究手段和工具, 综合分析处理基因

* 收稿日期: 2002-11-26

基金项目: 国家重点自然科学基金资助项目 (69933030)

作者简介: 王正华 (1962-), 男, 教授, 博士。

组图谱、基因序列等信息,阐明这些资料和数据所包含的生物学意义^[5,6]。目前的基因组信息学研究集中在功能基因组和比较基因组方面。

1.1 功能基因组信息学

功能基因组学的研究是在全基因组水平上对基因或其表达产物进行全面分析,目的是探究基因的时空差异表达情况。在研究层次上,它主要侧重于从基因的转录水平进行研究;在研究内容上,它包括基因功能发现、基因表达分析及突变检测;在分析手段上,目前主要使用基因表达的系统分析(serial analysis of gene expression, SAGE)、cDNA 微阵列(cDNA microarray)和 DNA 芯片(DNA chip)等分析技术。

1.1.1 基因表达的系统分析(SAGE)

SAGE 技术的主要理论依据是^[7]:来自 cDNA 3'端特定位置的一段序列(称为 SAGE 标签)能够区分基因组中 95% 的基因。通过对 cDNA 制备 SAGE 标签并将这些标签串联起来,然后对其进行测序,不仅可以显示各 SAGE 标签所代表的基因在特定组织中是否表达,还可以将各 SAGE 标签所出现的频率作为其所代表的基因表达丰度的指标^[8]。应用 SAGE 技术的一个必要前提是 GenBank 中必须有足够的某一物种的 DNA 序列资料,尤其是序列表达标签(EST)序列的资料。目前该技术在人类基因组研究中应用较为广泛^[9,10],主要侧重在对某些致病基因的研究上;但这种方法的缺点是不能够检测出稀有转录物。

1.1.2 cDNA 微阵列和 DNA 芯片技术

cDNA 微阵列和 DNA 芯片都是基于分子杂交的基因表达差异检测技术。二者的基本思路都是首先把 cDNA、EST 或基因特异的寡聚核苷酸固定在固相支持物上,并与不同来源的 cDNA 探针进行杂交,然后用特殊的检测系统对每个杂交点进行定量分析,从而反映出其所代表的基因在不同细胞、组织或器官中的相对表达丰度。这两项技术的优点是可以同时对大量基因,甚至整个基因组的基因的表达差异进行对比分析。cDNA 微阵列技术的主要优点是:灵敏度极高,十万分之一的低丰度仍可被检测出来;使用彩色荧光染料标记探针,在同一块阵列板上进行一次杂交实验就可以同时分析不同细胞间或不同环境下基因表达的差异。Desprez 等^[11]发展了利用尼龙膜作固相支持物和使用同位素标记探针进行杂交的 cDNA 表达阵列技术,从而降低了成本,但检测的灵敏度却降低了(即能检测到万分之一的丰度水平)。DNA 芯片技术具有高度并行化、多样化、微型化和自动化等特点,因而被广泛用于序列测定、转录分析以及基因诊断和药物设计等领域,成为功能组分析的支撑技术之一。

如何快速有效地对 DNA 芯片数据进行分析已经成为当前生物信息学的一项重要任务。聚类分析与模式识别是目前最常用的多元数据分析方法,近年来得到了广泛的研究^[12,13],但这种方法缺乏严格的统计检验基础,因而属于一种过渡性的方法。目前,仍需要发展新的分析方法与工具。近来,Li^[14]等人提出一种用遗传算法与 K-最优近邻的混合方法来实现分类,该方法是一种有监督的随机模式识别方法,可以进行显著性检验。Kaminski^[15]和 Nadon^[16]则从统计学角度分析了目前在大规模 DNA 芯片数据分析中常用的几种聚类方法,各自提出了一套分析流程,指出了目前分析方法中常见的几个误区。针对目前芯片数据分析中标准不统一的问题,Johnson 与 Lin^[17,18]等人发起了一个“微阵列数据分析评估准则”的会议,试图对不同方法在准确性与性能上进行比较,以促进相关研究的发展。

1.2 比较基因组学

比较基因组学的主要目的是通过模式生物基因组之间的比较与鉴别,为研究和理解生物的进化、人类遗传病候选基因的分离以及新的基因功能的预测提供重要依据^[19]。比较基因组学中主要使用各种分类方法和比对技术(序列比对,结构部件的比较等),这些方法也已渗透到蛋白质组信息学研究领域中,近年来该方向的研究主要侧重在对这些技术的改进提高上。序列比对广泛应用于生物信息学的各个研究中,如数据库搜索、进化发育分析、蛋白质的同源建模等。传统序列比对中的打分矩阵是恒定不变的,但实际的大分子序列在不同的位点其保守性并不相同,针对这个特点,最近 Muckstein 等人^[20]提出一种新的随机两序列比对方法,使得打分标准可以随着序列的不同而动态改变,具有更强的适应性。

在序列与结构相比对方面, Zimmer 等人^[21]利用蛋白质结构三维坐标数据的 Voronoi 分解,提出了一种新的“序列—结构”比对打分方法,在运行效率与折叠识别上有明显优势。

2 蛋白质组信息学的研究进展

2.1 蛋白质结构预测

在利用生物信息学进行蛋白质三维结构预测方面,总体上可分为同源建模、“从头预测”两大类方法^[22, 23]。

2.1.1 同源建模类方法

同源建模类方法是通过已知空间结构的蛋白质进行研究和分析,找出蛋白质序列与空间结构之间的联系,总结出一定的规律,建立一些经验规则,并将这些经验规则应用到待预测的靶序列上。同源建模类方法主要包括比较建模、折叠识别以及人工神经网络、隐 Markov 建模等方法。

比较建模(comparative modeling)主要依据是达尔文进化原理和“序列相似性必然蕴含着三维结构的相似性”这一经验性的原理^[24]。其代表性的研究工作有 Srinivasan^[23~25]等,比较建模的过程主要有以下几个步骤:(1)搜索模板序列。即在已知结构的蛋白质数据库中搜索与靶序列足够相似的序列,如果找不到这样的模板,则应当选择使用其它的建模方法。(2)对靶序列与模板序列进行多重比对。目的是在允许出现插入与删除的情形下,使靶序列的每个残基与模板序列各个残基获得最大程度的匹配。(3)从模板的空间结构数据导出靶序列的空间结构信息。在建立好的多重比对中,将模板序列中的结构信息复制给与之相匹配的靶序列残基,这样就得到了靶序列的主链结构信息。(4)建立靶序列的侧链结构及回环区域模型。蛋白质侧链构象的预测通常是靠经验性数据来辅助完成的,而回环结构信息可通过统计其它蛋白质的相关信息或是用“全新设计”的办法进行。(5)进一步精化模型。由于靶序列与模板序列的差异性,上述得到的靶蛋白结构信息可能存在许多不合理的地方(比如,多个原子之间会出现位置重叠的情况),尚需要作适当的修正,有许多方法可以实现这样的优化^[26]。

折叠识别(fold recognition)方法适用于模板蛋白与靶蛋白序列上不太相似但结构上却相似的情况。据估计,有50%以上的基因组蛋白与已知结构的蛋白质在空间结构的折叠方面有足够的相似度^[27]。由于不能用单纯的序列比较方法获得这些模板蛋白,必须借助于靶序列与模板结构相比较的办法,穿线法(threading)是其中较为成功的一种^[28],其技术要点包括:(1)必须有一个已知三维折叠结构的数据库。这个数据库的代表性的三维折叠结构信息可以用于“序列—结构”之间的比较。(2)执行“序列—结构”的比较。用靶序列到上述三维折叠结构数据库中搜索,以找到相似的折叠结构,这需要“序列—结构”之间的比较,这种比较一般要考虑到特定折叠结构中氨基酸的偏好性^[29]。(3)“序列—结构”中的优化技术和统计显著性评估方法。当考虑空位的插入与删除情形时,靶序列与三维折叠结构相比较是一个 NPC 问题,必须使用近似或启发式的方法。每个靶序列与三维折叠结构库比较都会得到一个打分数值,但这个数值是否真正具有生物学上的意义,则需要进行统计学的显著性检验^[30]。

在神经网络(ANN)与隐 Markov 建模(HMM)方法中,一般是利用数理手段对蛋白质进行建模,并通过已有信息的学习来获得蛋白质序列与结构之间关系的相关知识,然后将这些知识应用到靶蛋白质的结构预测中。近年来这方面的研究比较多,如 Krogh^[31], Karplus^[32], Trentin^[33], Baker^[34], Baldi^[35]等。在隐 Markov 建模方法中,需要构建一阶有限状态 Markov 链,并将蛋白质序列中的各个氨基酸残基视为这些状态的输出内容,通过对特定样本的学习,将获得的关于序列与结构联系方面的知识“存储”在模型的各个参数中,最后使用这个学习好的模型去预测靶蛋白质的结构。神经网络模型也有相似的原理,区别在于它将蛋白质序列特征作为网络模型的输入、将对应的结构类型作为网络模型的输出。最近, Cai 等人^[36]使用支持向量机网络模型对蛋白质二级结构分类,提高了分类效率,同时也较好地避免了在一般的神经网络模型中常见的过适应(overfitting)的缺点。在模型本身的研究方面, Kundu^[37]和 Wessels 等^[38]将 HMM 易于训练的特性与 ANN 的强识别能力相结合,构建了一个神经网络与 HMM 的混合模型。在通常的 HMM 中都使用了无后效性的假设,为了避免由此所带来的问题, Bengio

等人^[39]在数理推导的基础上,建立了“正一反”双向的隐 Markov 模型,新模型的类区分能力更强,并具有很好的自适应特性。

2.1.2 “从头预测”类方法

“从头预测(ab initio)”方法不依赖于已有的结构数据信息,直接从蛋白质序列利用分子动力学原理预测和推断结构信息。这类方法主要依据了一个基本热力学假定:一个蛋白质分子的溶液中的天然构象对应于热力学上最稳定、自由能最低的构象。Bonneau^[40]和 Fischer^[23]等人的工作是这方面的代表性研究,其关键的技术主要包括:(1)收集有各种蛋白质几何外形的数据库。考虑到后续步骤的复杂性,各蛋白质外形数据都有某种程度的近似与简化(比如,忽略掉体积小的原子,简化侧链结构等)。(2)一个能量评估函数。通常基于统计学与分子动力学原理构造而成,主要用于评估蛋白质在不同构象下的自由能值^[41]。(3)对各种蛋白质构象的搜索技术。大多数从头预测方法使用 Monte Carlo、模拟退火和遗传算法等技术对蛋白质构象空间进行搜索,并对每个构象进行能量评估,以找到自由能最低的构象。

2.2 分子模拟与药物设计

传统的药物研制主要是从大量的天然产物、合成化合物以及矿物中进行筛选,得到一个可供临床使用的药物要耗费大量的时间与金钱。近年来,相当数量的蛋白质以及一些核酸、糖类三维结构已被人们精确测定,使得借助生物信息学进行药物设计成为可能。原则上,任何有功能的蛋白质都可以作为蛋白质工程的改造对象,但实际上选择目标时往往要考虑多个方面的因素^[42]:改造对象有没有测定空间结构;结构和生物功能的联系是否明确;所选对象的重要性;是否易于进行分子设计和最后的基因工程生产等等。当前的分子设计主要以能显示图形图像的计算机为工具,在了解需改造蛋白质的性能及其结构的基础上,依据分子动力学原理以及蛋白质分子结构构建相应的模型,提出蛋白质改性的设计方案。

利用生物信息学对蛋白质分子进行理论模拟与结构预测为天然生物大分子的改性和基于受体结构的药物分子设计提供了依据^[37]。一般情况下,基于受体结构的药物筛选是药物设计的首要环节。药物的治疗作用主要是通过药物与受体的相互作用来完成的,有很多生物大分子可以作为药物设计的受体模型。当前这方面的研究十分活跃^[43],例如:Rich^[44]等人基于酶分子结构的药物设计研究,Waszkowycz^[45]基于抗体结构的药物设计研究,Klabunde^[46]等人基于细胞表面受体结构的药物设计研究等。同时,以结构为基础的药物设计和分子模拟也是密不可分的。在 Lin^[47]和 Zhu^[48]等的研究中总结了这种方法的一般过程:首先,在含有大量化合物三维结构的数据库中,搜索能与生物大分子靶点匹配的化合物,或者搜索能与结合药效团相符的化合物,利用这种方法,可以快速地确定具有假定药效团的化合物。其次,在对先导化合物与靶点的嵌合结构作了分子模拟研究后,可通过优化先导化合物进一步提高其与靶点的匹配性,通常将优化选择在分子与活剂接触部分的效价中性部位或分子中与靶点结合不重要的部位。

3 未来生物信息学展望

综观当前后基因组时代的研究现状与进展,可以看到生物信息学的研究呈以下几个趋势:

(1) 研究目标由“组成”转向“功能”

以往的分析大多是通过同源性搜索、模式发现、多序列比对以及序列聚类分析等比较技术,来实现对序列组成与一级结构的理解,这些方法的共同特点是过分依赖于生物大分子序列组成。然而,为了理解不同生物大分子的功能差异以及同一生物大分子在时间跨度上的变化情况,就必须对其功能表达情况作深入的分析与研究。可以预见,近年来广泛应用的基因微阵列分析技术和 NMR 分子识别技术将会发挥越来越重要的作用。

(2) 研究内容由“静态”转向“动态”

为进一步理解细胞信号的规律、掌握各种大分子的代谢途径、揭示生命的奥秘,仅对静态的单个生物大分子进行研究是不够的,必须研究基因表达过程的动态特性。但由于基因表达的复杂性,目前的研究只能通过分子模拟的手段来进行。尽管如此,有些研究者已经开始使用一些数理模型对基因调控网络进行研究。

(3) 研究角度由“局部性”转向“整体性”

由于数据的不完整或分析软件处理结果的差异,以往的研究分析只能自始至终使用同一个软件工具,研究分析的对象只能集中在一个局部的数据集上。目前随着已完成基因组测序的生物物种数目的增加,以及分析工具的日益丰富,未来的研究会充分利用比较基因组学的分析方法,对各个物种的基因组信息进行综合分析比较,最终得到整体性的生物学结论。

(4) 研究方法由“单一”转向“综合”

传统的研究方法无论在广度还是深度方面都有诸多限制。如在广度方面,主要采用“每次只研究一个基因”的办法;而在深度方面,则使用多组实验分析比较时“每次只修改一个变元”的办法。这些办法对于以往实验数据较少的情况是适用的,但对于高通量表达的数据分析以及基因组水平的数据分析,则无能为力了。只有综合使用各种数理统计方法和信息分析处理技术,才能满足需求。

4 结论

作为计算机科学和分子生物学等形成的交叉科学,生物信息学已经在多个领域的研究中取得了可观的成果,但就其整个学科而言,其发展历史不长,因此无论是理论上还是方法上,都还有许多亟待解决的问题,蕴藏着极大的研究潜力。

针对目前的研究现状,我们认为生物信息学尚需解决以下几方面的问题。

4.1 加强基础科学研究

生物信息学对分子生物学、统计生物学、计算机科学和应用数学等相关学科的发展提出了挑战,同时也为这些学科的发展提供了新素材、新工具和新思路。如利用比较基因组学进行分子水平的系统发育分析,研究生命是从哪里起源的,生命是如何进化的,遗传密码是如何起源的,估计最小独立生活的生物至少需要多少基因,这些基因是如何使它们生活起来的等生命科学的基本问题,将为探索物种起源、揭示生命奥秘提供有力的帮助。

4.2 研发高效算法及支持软件

生物信息学中多处存在着算法优化及建模问题。基因测序过程中的序列拼接与组装、图谱绘制、序列比对、系统发育分析、蛋白质结构预测、分子模拟及药物设计等都需要高效算法及易用软件的支持。例如:(1)在编码区域和基因识别方面,需要发展更精确的识别算法;(2)在蛋白质结构预测方面,一方面为了进一步提高预测的准确率,需要现有的理论分析方法,综合使用统计方法、隐 Markov 模型、神经网络方法以及支持向量机等;另一方面,在预测效果的评估方面,需要发展一系列自动评估工具;(3)在序列拼接与组装过程中,充分挖掘现有计算资源,发展高性能计算技术;(4)编制具有良好人机接口的、支持分布式资源的可重用软件系统。

4.3 发展数据库集成与资源再加工技术

网络上的公共数据库在物理上散布于各地,数据资源都带有不同程度的信息冗余或缺失,需要在数据加工基础上建立二级以上的数据库及相应的处理分析工具,以便为世界各地的科学家提供快速高效的信息服务。相应地,应当在数据质量控制、数据资源的标准化、网络传输、多数据库系统集成以及数据挖掘、知识发现等各方面加以研究,发展相应的技术。

参考文献:

- [1] Baldi P, Brunak S. Bioinformatics: The Machine Learning Approach[M]. Cambridge, Mass.: MIT Press, 2001.
- [2] 欧阳曙光, 贺福初. 生物信息学: 生物实验数据和计算技术结合的新领域[J]. 科学通报, 1999, 44(14): 1457-1468.
- [3] 陈润生. 当前生物信息学的重要研究任务[J]. 生物工程进展, 1999, 19(4): 11-14.
- [4] Yockey H P, Platzman R P, Quastler H. Symposium on Information Theory in Biology[C]. Pergamon Press, New York, London, 1958.
- [5] Smith T F. Functional Genomics-bioinformatics Is Ready for the Challenge[J]. Trends in Genetics, 1998, 14: 291-293.
- [6] Hill D P, Davis A P, Richardson J E, et al. Strategies for Biological Annotation of Mammalian Systems: Implementing Gene Ontologies

- in Mouse Genome Informatics[J]. *Genomics*, 2002, 74: 121– 128.
- [7] Yamamoto M, Wakatsuki T, Hada A, Ryo A. Use of Serial Analysis of Gene Expression (SAGE) Technology[J]. *Journal of Immunological Methods*, 2001, 250(1– 2): 45– 66.
- [8] Scott H S, Chrast R. Global Transcript Expression Profiling by Serial Analysis of Gene Expression (SAGE) [J]. *Genetic Engineering*, 2001, 23: 201– 219.
- [9] Tan S S, Gunnensen J, Job C. Global Gene Expression Analysis of Developing Neocortex Using SAGE[J]. *The International Journal of Developmental Biology*, 2002, 46(4): 653– 660.
- [10] Ye S Q, Lavoie T, Usher D C, Zhang L Q. Microarray, SAGE and Their Applications to Cardiovascular Diseases[J]. *Cell Research*, 2002, 12(2): 105– 115.
- [11] Desprez T, Amsellem J, et al. Differential Gene Expression in Arabidopsis Monitored Using cDNA Arrays[J]. *The Plant Journal*, 1998, 14(5): 643– 652.
- [12] McLachlan G J, Bean R W, Peel D. A Mixture Model-based Approach to the Clustering of Microarray Expression Data[J]. *Bioinformatics*, 2002, 18: 413– 422.
- [13] Masys D R, Welsh J B, et al. Use of Keyword Hierarchies to Interpret Gene Expression Patterns[J]. *Bioinformatics*, 2001, 17: 319– 326.
- [14] Li L, Clarice R, et al. Gene Selection for Sample Classification Based on Gene Expression Data: Study of Sensitivity to Choice of Parameters of the GA/KNN Method[J]. *Bioinformatics*, 2001, 17: 1131– 1142.
- [15] Kaminski N, Friedman N. Practical Approaches to Analyzing Results of Microarray Experiments[J]. *American Journal of Respiratory Cell and Molecular Biology*, 2002, 27(2): 125– 132.
- [16] Nadon R, Shoemaker J. Statistical Issues with Microarrays: Processing and Analysis[J]. *Trends in Genetics*, 2002, 18(5): 265– 271.
- [17] Johnson K F, Lin S M. Critical Assessment of Microarray Data Analysis: the 2001 Challenge[J]. *Bioinformatics*, 2001, 17: 857– 858.
- [18] Johnson K F, Lin S M. Microarray Analysis: A Comparative Approach[J]. *Molecular Cancer Therapeutics*, 2002, 1(3): 237– 238.
- [19] Califano A. Advances in Sequence Analysis[J]. *Current Opinion in Structural Biology*, 2002, 11: 330– 333.
- [20] Muckstein U, Hofacker I L, Stadler P F. Stochastic Pairwise Alignments[J]. *Bioinformatics*, 2002, 18: 153S– 160S.
- [21] Zimmer R, Wohler M, Thiele R. New Scoring Schemes for Protein Fold Recognition Based on Voronoi Contacts[J]. *Bioinformatics*, 1998, 14: 295– 308.
- [22] Baker D, Sali A. Protein Structure Prediction and Structural Genomics[J]. *Science*, 2001, 294(5540): 93– 96.
- [23] Sieu N, Fischer D. Convergent Evolution of Protein Structure Prediction and Computer Chess Tournaments: CASP, Kasparov, and CAFASP[J]. *IBM Systems Journal*, 2001, 40(2): 410– 425.
- [24] Sali A. Modeling Mutations and Homologous Proteins[J]. *Current Opinion in Biotechnology*, 1995, 6(4): 437– 451.
- [25] Srinivasan N, Guruprasad K, Blundell T L. Comparative Modelling of Proteins, Protein Structure Prediction: A Practical Approach[M]. M. J. E. Stemberg, Editor, IRL Press at Oxford University Press, Oxford, 1996: 111– 140.
- [26] Durbin R, Eddy S R, et al. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*[M]. Cambridge University Press, 1998.
- [27] Fischer D, Eisenberg D. Assigning Folds to the Proteins Encoded By the Genome of *Mycoplasma Genitalium*[J]. *Proceedings of the National Academy of Sciences(USA)*, 1997, 94: 11 929– 11 934.
- [28] Xu D, Unseren M A, Xu Y, Ueberbacher E C. Sequence-structure Specificity of a Knowledge Based Energy Function at the Secondary Structure Level[J]. *Bioinformatics*, 2000, 16: 257– 268.
- [29] Jones D T. Protein Structure Prediction in the Post-genomic Era[J]. *Current Opinion in Structural Biology*, 2000, 10: 371– 379.
- [30] Bryant S H, Altschul S F. Statistics of Sequence-structure Threading[J]. *Current Opinion in Structural Biology*, 1995, 5: 236– 244.
- [31] Krogh A, Riis S K. Hidden Neural Networks[J]. *Neural Computation*, 1999, 11(2): 541– 563.
- [32] Karplus K, Sjölander K, et al. Predicting Protein Structure Using Hidden Markov Models[R]. UCSG-CRL-97-13, 1997.
- [33] Trentin E, Gori M. A Survey of Hybrid ANN/HMM Models for Automatic Speech Recognition[J]. *Neurocomputing*, 2001, 37: 91– 126.
- [34] Bystroff C, Thorsson V, Baker D. HMMSTR: A Hidden Markov Model for Local Sequence-structure Correlations in Proteins[J]. *Journal of Molecular Biology*, 2000, 301: 173– 190.
- [35] Baldi P, Brunak S, et al. Bidirectional Dynamics for Protein Secondary Structure Prediction[J]. *Lecture Notes in Computer Science*, 2001, 1828: 80– 86.
- [36] Cai Y D, Liu X J, et al. Support Vector Machines for the Classification and Prediction of Beta-turn Types[J]. *Journal of Peptide Science*, 2002, 8: 297– 301.
- [37] Kundu A, Bayya A. Speech Recognition Using Hybrid Hidden Markov Model and NN Classifier[J]. *International Journal of Speech Technology*, 1998, 2: 227– 240.
- [38] Wessels T, Omlin C W. Refining Hidden Markov Models with Recurrent Neural Networks[C]. *South African Telecommunications Networks and Applications Conference (SATNAC)*, 1999.

估参数。

(3) 三维动画显示模块。以三维动画方式显示空战过程中红蓝双方导弹发射脱靶与遇靶过程,并提供着火及爆炸的画面和音响。

2.2.4 人机交互软件

人机交互模块为指挥人员提供直观、方便和灵活的操作界面。它支持所有显示要素的用户化定义,为仿真模型的交互式控制提供输入接口,同时还允许用户在实际仿真过程中灵活控制或调整显示区域。

3 结论

实际应用表明:本文提出的多机空战评估系统设计先进、功能完整,基本上满足多机空战对抗训练与指挥的要求,为科学、客观地评估空战结果提供了一种新的技术途径。考虑到目前系统存在的不足,今后多机空战对抗评估系统发展主要解决下面几个问题:一是提高飞机实时信息获取量,如增加向地面实时传送飞机姿态信息和火控系统的开关信息,从而增加仿真模型逼真性;二是尽可能采用三维立体显示,更真实反映红蓝双方空中立体态势,增强实战感;三是考虑电子干扰与反干扰的仿真模型,提高评估系统实战适应性。

参考文献:

- [1] 杨乐平,等.多机空战对抗评估系统总体设计报告,国防科技大学内部技术报告[R],1999.
- [2] Finell F. Analysis of a Generic Air-to-air Missile Simulation Model[R], N94-36438, 1994.
- [3] 季节,许云剑,等.世界机载雷达手册[M].北京:航空工业出版社,1989.
- [4] 熊光燧,彭毅,等.先进仿真技术与仿真环境[M].北京:国防工业出版社,1997.

(上接第6页)

- [39] Bengio Y, Frasconi P. Input-Output HMMs for Sequence Processing[J]. IEEE Transactions on Neural Networks, 1996, 7(5): 1231-1249.
- [40] Bonneau R, Tsai J, et al. Rosetta in CASP4: Progress in Ab Initio Protein structure prediction[J]. Proteins: Structure, Function and Genetics (available also from <http://depts.washington.edu/bakerpg/papers/Bonneau-Proteins.pdf>), 2001.
- [41] Simons K T, Bonneau R, Ruczinski I, Baker D. Ab Initio Protein Structure Predictions of CASP 3 Targets Using ROSETTA[J]. Proteins: Structure, Function, and Genetics Supplement, 1999, 3: 171-176.
- [42] 赵善荣,林茂伟,陈凯先.生物信息学在药物设计中的应用[J].药学进展,1997,21(2):65-70.
- [43] Carlson H A. Protein Flexibility Is an Important Component of Structure-based Drug Discovery[J]. Curr Pharm Des, 2002, 8(17): 1571-1578.
- [44] Rich D H, Bursavich M G, Estiarte M A. Discovery of Nonpeptide, Peptidomimetic Peptidase Inhibitors that Target Alternate Enzyme Active Site Conformations[J]. Biopolymers, 2002, 66(2): 115-125.
- [45] Waszkowycz B. Structure-based Approaches to Drug Design and Virtual Screening[J]. Curr Opin Drug Discov Devel, 2002, 5(3): 407-413.
- [46] Klabunde T, Hessler G. Drug Design Strategies for Targeting G-protein-coupled Receptors[J]. Chembiochem, 2002, 3(10): 928-944.
- [47] Lin J H, Perryman A L, Schames J R, McCammon J A. Computational Drug Design Accommodating Receptor Flexibility: the Relaxed Complex Scheme[J]. J Am Chem Soc, 2002, 124(20): 5632-5633.
- [48] Zhu J, Fan H, Liu H, Shi Y. Structure-based Ligand Design for Flexible Proteins: Application of New F-DycoBlock[J]. J Comput Aided Mol Des, 2001, 15(11): 979-996.