

SARS 病毒抗原表位预测

李伍举 刘涛 范明

军事医学科学院基础医学研究所，北京：100850

摘要

[目的] 采用集 Hopp&Woods 亲水性、Janin 表面可及性、Karplus-Schulz 主链柔软性和电荷分布为一体的综合性抗原表位预测方法和蛋白质二级结构预测对 SARS 病毒的两个膜蛋白 S 和 M 进行抗原表位预测，以便为 SARS 病毒的疫苗设计提供依据。[结果]通过运用 Goldkey 等软件分析了 SARS 病毒的两个膜蛋白 S 和 M 的抗原表位，分别获得了 14 个和 7 个可能的抗原表位。

一、前言

目前，SARS 病毒对人类的危害众所周知，我国政府已投入巨大的人力、财力和物力向 SARS 宣战，特别是军事医学科学院几乎与世界同步公布了 SARS 病毒的全基因组序列，为 SARS 的诊断提供了强有力的基础。从生物信息学的角度来看，此病毒的全基因组序列蕴藏着决定 SARS 病毒表型如致病性、抗原性等特征，如何利用生物信息学手段，从 SARS 病毒的全基因组序列中挖掘尽可能多的信息，为 SARS 病毒的诊断及疫苗的研制提供服务，已成为我们生物信息学工作者的当务之急，为此，根据 SARS 病毒的特点，重点选择了位于表面的两个膜蛋白 S 和 M¹，利用我们以前开发的 DNA 和蛋白质序列分析软件 Goldkey²，采用集 Hopp&Woods 亲水性、Janin 表面可及性、Karplus-Schulz 主链柔软性、电荷分布、综合性抗原表位预测方法³和蛋白质二级结构预测对 SARS 病毒的 S 蛋白、M 蛋白进行抗原表位预测，以便为 SARS 病毒的疫苗设计提供依据，本文就是对这一工作的系统总结。

二、材料与方法

2.1、基因组序列数据

分别选取了 9 个 SARS 病毒的全基因组序列，其中美国一株、加拿大一株、中国香港 2 株、中国北京 4 株和中国广州 1 株。

2.2、软件

为了预测 SARS 病毒的抗原表位，主要采用了两个软件，它们分别是 ClustalX⁴ 和 Goldkey，基于 ClustalX 进行基因组和蛋白质的多序列比较，基于 Goldkey 预测 SARS 病毒两个膜蛋白 S 和 M 的抗原表位。

三、结果与讨论

3.1、SARS 病毒的基因组比较分析

从 GenBank 数据库中，将 9 个 SARS 病毒的基因组序列下载至本地机上，按多序列比较程序 ClustalX 的格式要求进行基因组数据整理，然后进行多序列比较，结果表明，这些基因组序列极其保守，变异甚微，从而说明可以其中一株为基础进行抗原表位预测与疫苗设计，为此，我们以加拿大 SARS 病毒株为基础进行抗原表位预测，并通过对 SARS 病毒的 S 和 M 蛋白的进一步比较表明，这些重要蛋白非常保守。

3.2、SARS 病毒的 S 和 M 蛋白的抗原表位综合预测

首先，运用我们自行开发的 Goldkey 软件读入 SARS 病毒的 S 和 M 蛋白质序列，然后，利用我们提出的抗原表位综合预测方法，分别预测 S 和 M 蛋白的抗原表位，结果如图 1 和图 2 所示。根据抗原表位的综合预测方法，其综合指标处于峰值位置的片断有可能为抗原表位，据此，预测出的 SARS 病毒 S 蛋白的可能抗原表位为：S12-T20, P28-M37, A90-N119, L170-F187, K265-A275, Y338-C348, V389-P399, N427-N457, G536-G570, S664-Y677, G751-K768, Q895-K911, V1047-A1062 和 N1116-I1151。SARS 病毒 M 蛋白的可能抗原表位为：E10-Q18, F36-L45, A103-I117, R130-E136, C158-I167, S172-Y178 和 S183-A194。另外，我们还对这两个氨基酸序列进行了亲水性等单指标分析，结果表明：单指标分析结果与上述综合指标分析结果有较大的一致性，从而从生物信息学的角度，更加证实了上述预测的结果。

3.4、SARS 病毒的 S 和 M 蛋白的二级结构预测

作为蛋白质的可能抗原位点，要求其必须位于蛋白质空间结构的表面，所以

必须对其进行空间结构预测,才能较好地揭示其可能抗原位置,但是,通过扫描已知结构的蛋白质数据库表明,目前还不存在与 SARS 病毒的两个膜蛋白 S 和 M 的同源蛋白,因此,目前不能预测其空间结构。为此,我们对这两个重要蛋白进行了二级结构预测,结果如图 3 和图 4 所示,以便有助于分子生物学工作者从上述可能的抗原表位列表中做出进一步选择。

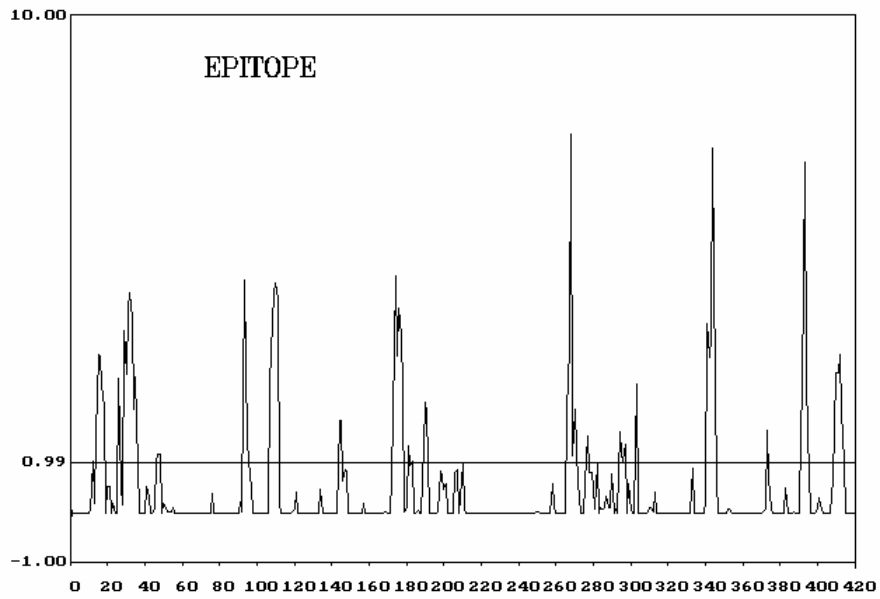
四、总结与体会

本文总结了利用我们自行开发的 Goldkey 软件对 SARS 病毒 S 和 M 蛋白的抗原表位预测情况,基于多个参数找出了 S 和 M 蛋白的可能抗原表位,从而为疫苗设计等相关实验提供依据,毫无疑问,提高抗原表位预测的准确性将加速有关疫苗的设计进程,早日造福人类。目前,已有许多抗原表位预测方法和程序,为此,我们进一步开展两方面工作,一方面,利用网上免费抗原表位预测程序,对 SARS 病毒的 S 蛋白与 M 蛋白进行抗原表位预测,以便进行比较,确定比较好的可能抗原表位,以便用于疫苗设计;其次,开展抗原表位预测方法研究,提高预测精度。通过这项工作的开展,我们的体会是,生物信息学是一门非常有应用前景的学科,基于生物学实验(Wet Lab.)产生生物信息,基于生物信息开展生物信息学研究(Dry Lab.),从而实现从生物信息(information)获取到知识发现(Knowledge discovery)的过程,然后利用发现的知识来解决生物学问题,这是一个辩证循环不断上升且相互促进的过程,其共同目标都是为了更好地解决生物学问题,使人们的生活更加美好。

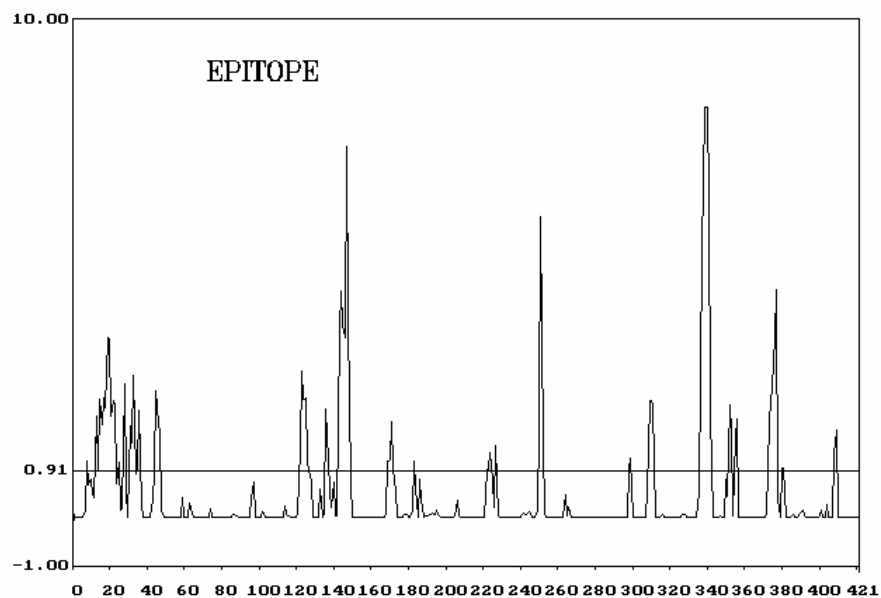
五、参考文献

- 1、Marra, MA et al. The Genome sequence of the SARS-Associated Coronavirus, Science, Published online May 1, 2003; 10.1126/science.1085953 (Science Express Research Articles)
- 2、吴加金等,Goldkey:核酸与蛋白质序列分析的软件系统,生物技术通讯,1994, 5: 189-193。
- 3、万涛等,蛋白顺序性抗原决定簇的多参数预测,中国免疫学杂志,1997, 13: 329-333。

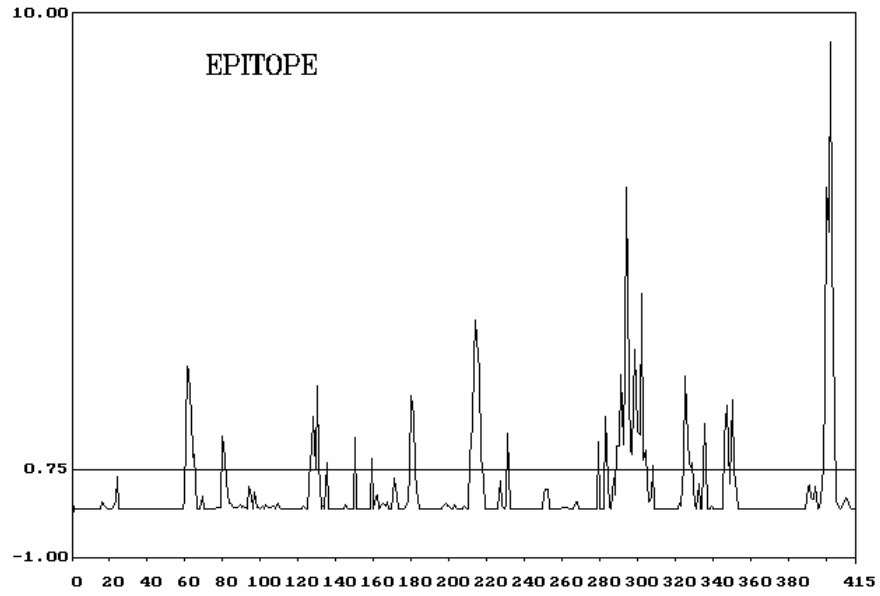
- 4、 Thompson,JD et al. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools , Nucleic Acids Research, 1997, 24:4876-4882.



epitope analysis of NP_828851_1_420

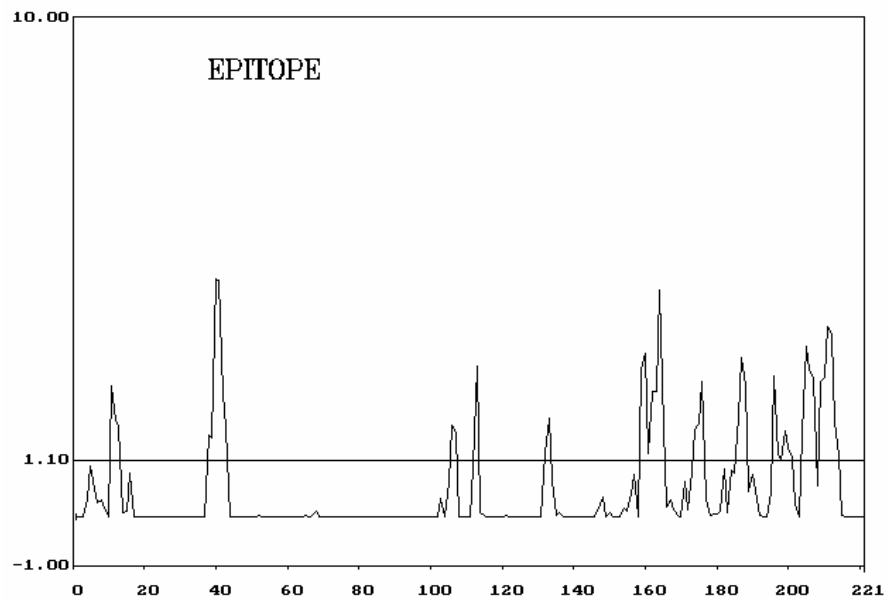


epitope analysis of NP_828851_420_840



epitope analysis of NP_828851_840_1255

图 1：SARS 病毒 S 蛋白的抗原表位的综合预测，该序列全长为 1255 个氨基酸，为了使显示的氨基酸序号比较清楚，整个序列被分为 3 段，分别进行分析。



epitope analysis of NP_828855

图 2：SARS 病毒 M 蛋白的抗原表位的综合预测，该序列全长为 221 个氨基酸。

< or > =helix,E =extended,T =turn,- =coil

1 MFIFLLFLTLTSGSDLRCTTFDDVQAPNYTQHTSSMRGVVYPDEIFRSDTLYLTLQDLFL
HHHHHHHCCTTTTCCCTEETTTTTTCCCTTTTCCCTTTTCCCTEEEEETTTCTEEEEEE

61 PFYSNVTGFHTINHTFGNPVIFPKDGIYFAATEKSNVVRGWVFGSTMNKSQSVIIINNS
TTTCTCTTCCCTTTTCCCEETETCCEHHHHHTCCEEEEECCCCCTTCCCEEEETCC

121 TNVVI RACNFELCDNPFVAVSKPMGTQHTMIFDNAFNCTFEYISDAFSLDVSEKSGNFK
CCEEEEEETTTTTTCTHHHHCCCTCCEEEEEETTTTTCTTETHTHHHHHHCTTTTHH

181 HLREFVFKNKDGLYVYKGYQPIDVVRDLPSGFNTLKP I FKLPLGINITNFRAILTAFSP
HHHHHHHTTTTTEEEEETTTTTEEEEEEECCTTTCCCTTEEECTCEEEEEEEEECCCC

241 AQDIWGTSAAAYFVGYLKPFTFMLKYDENGITDAVDCSQNPLAELKCSVKSEIDKGIY
TTTTTCCCCCEEEETCCCTEEEEHCTTTTCCCEEEETTTCCCHHHHHHHHHHHHTTTT

301 QTSNFRVPSGDVVRFPNITNLCPFGEVFNATKFPSVYAWERKKISNCVADYSVLYNSTF
TCCTEEECTCTEEEEETTTTCTTTEHHHHCCCCCHHHHTTTTTEEEEETEEETTCTE

361 FSTFKCYGVSATKLNLCFSNVYADSFVVKGDDVRQIAPGQTGVIADYNYKLPDDFMGCV
TTTTTTTTCEETTTTEEEEEETTHHEEETTHHEEECTTCEEEETTTCTTTEEEEE

421 LAWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCPPALNCYWPLND
HHHCCTTCTCCCTTTTTTTEEEETTTCCCTHTTTCCCTCTTTCTTEETTETTTETTTT

481 YGFYTTTGI GYQPYRVVLSFELLNAPATVCGPKLSTDLIKQCVNFNFNGLTGTGVLTP
TTTEETTTTCCCEEEEEEEHHHCCHTHEEETTCEEEETCTEEETTTTTTCCCEEEEC

541 SSKRFQPFQFGRDVSDFDTSVRDPKTSEILDISPCAFGGVSVITPGTNASSEVAVLYQD
CTTTTCTCTTTTTTCTTTTCTTCTCCEEEETTTTTTTEEEECTCCCCCEHEHEETT

601 VNCTDVSTAIHADQLTPAWRIYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASY
TEETTCHHHHTTCCCTTEETTTTCCCEEEETECCHHHCTTTTTTEETTTTEEEEE

661 HTVSLLRSTSQKSI VAYTMSLGADSSI AYSNNTIAIPTNFSISITTEVMPVSMAKTSVDC
EEEEETCTCTCEEEEEEECCCCCEETCTCEEEETTTCCCECEEEHHHHHHHCETT

721 NMYICGDSTECANLLLQYGSFCTQLNRALSGIAEQDRNTREVFQVQMYKTPTLKYFG
TEETTTTCHHHHHHTTTTTEEECTTCCCEHHHTTCCCHHHHHHTTECCTEEETT

781 GFNFSQILPDPLKPTKRSFIEDLLFNKVTLADAGFMKQYGECLGDI NARDL I CAQKFNGL
TCCTCTCCCTCCCTCEEEHHHHHHHHHHHHHTTTTTTTTCHHHHHHHHTTTTT

841 TVLPPLLTDDMI AAYTAALVSGTATAGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYE
 CEEEEHHHHHHHHHHHECCCCCTCCCCCCHHEHHHHHHHTTTTCCCEEEEECH

901 NQKQIANQFNKAISQIQESLTTTSTALGKLQDVVNQNAQALNTLVKQLSSNFGAIVSSVLN
 HHHHEHCHTCCCCHEEECCCCCHHCHEEEEEHCCHHHHEEEEECCTTCCCEEEEE

961 DILSRDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSK
 TEHHHHHHHHHHHHHEEETCCCEEEEEEEEEHHHHHHHHHHHHHEEEEEHHHT

1021 RVDFCGKGYHLMSFPAAPHGVVFLHVTVPSQERNFTTAPACHEGKAYFPREGVVFVN
 TEEETTTTTEEEETTHHTTTHEEEEEECTTTTTTTTCHTEEHTTTTTTEHTTTEETT

1081 GTSWFITQRNFFSPQITTDNTFVSGNCDVVIGIINNTVYDPLQPELDSFKEELDKYFKN
 TTTTEECTCTCCTTEETETTTTTTTTTTEEEHHHTTTTTTEHHHHHHHHHHHHHTT

1141 HTSPDVLGDISGINASVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYVWL
 TTCTTTTETHHEEEEEEEEEEECHHHHTCCTEEHHHHHCCCTTEETCCCTEEEE

1201 GFIAGLIAIVMVTILLCCMTSCCCLKGACSCGSCCKFDEDDSEPVLKGVKLHYT
 TTHHHHHHEEEEEHTTE

图3 : SARS病毒S蛋白NP-828851二级结构预测结果

< or > =helix, E =extended, T =turn, - =coil

1 MADNGTITVEELKQLLEQWNLVIGFLFLAWIMLLQFAYSNNRNFYI IKLVFLWLLWPVT
 HHTCCCCHHHHHHHHHHEEEEEHHHHHHHTCTCTTTTEEEEEEEETCTCCTC

61 LACFVLAAYRINWVTGGIAIAMACIVGLMWLSYFVASFRLFARTRSMWSFNPETNILLN
 EEEHHHHHEEEETCCCCHEHHEEEEEHEEEEEHHHHHHCTTTTTCCCTTTEEEE

121 VPLRGTIVTRPLMESELVIGAVIRGHLRMAGHSLGRCDIKDLPKEITVATSRTLSTYKYL
 ECTTEEEEEECCHHHHHHHHEEETCCCTTTTCTTTTETCCHHHHEEEECTTEEEETC

181 GASQRVGTDSGFAAYNRYSRIGNYKLNTDHAGSNDNIALLVQ
 TEEEEETCCCCEEEETTTTTTTC

图4 : SARS 病毒 N 蛋白 NP-828855 二级结构预测的文字表示