

## 人工智能如何助力开发癌症疫苗

人类免疫系统是对抗各种疾病的利器。通过模拟病毒特征，可以定向刺激免疫系统，让我们的身体产生抗体，从而将入侵者其消灭，这就是疫苗的基本原理。例如，经过科学家们的努力，针对新冠病毒的疫苗已经陆续上市。

既然疫苗如此有效，能否开发出干掉癌症细胞的疫苗呢？答案是“可以的”。早在一百多年前就有科学家进行了相关研究，且已经有一些针对特定癌症的疫苗投入临床应用，例如作为预防性的人乳头状瘤病毒（HPV）疫苗和作为治疗性的前列腺癌疫苗（Sipuleucel-T）。然而，到目前为止，这种免疫疗法还没有大规模应用，这是为什么呢？

我们知道开发疫苗最关键的是找到坏蛋们（如癌细胞）的标志性特征，这些特征称为**抗原**。那么，直接把癌症细胞揪出来，依葫芦画瓢造个疫苗不就好了？然而，癌细胞非常狡猾，把自己伪装成正常细胞，连我们的免疫系统都很难发现他们。因此，以往大多数癌症疫苗只能基于**肿瘤相关抗原**，而不是**肿瘤特异性抗原**。这种相关性抗原确实存在于癌细胞中，但很多时候也存在于正常细胞中，只不过相对较少。因此，基于这种抗原激发出来的免疫反应在消灭癌细胞的同时也会无情地消灭正常细胞。所谓“杀敌一千，自损八百”。

这就激发出科学家们的**好胜心**：人类已经能把整个细胞揪出来进行基因测序了，难道还不能找到一个肿瘤特异的抗原吗？于是，科学家们开始将癌细胞和正常细胞放在一起“找茬”，检测出癌细胞中发生变异的蛋白质片段，再找到那些可以激活免疫反应的变异片段，那么我们的抗癌疫苗就有希望了【1】。这些能激发免疫反应的蛋白质片段称为**新生抗原（Neoantigen）**。

然而，道路还是曲折的。科学家们发现，癌症和癌症之间的新生抗原有所不同，这意味着针对不同癌症要设计不同的疫苗。这个倒不是大问题。糟糕的是，即使是同一种癌症，不同患者的新生抗原也是不同的。这就麻烦了，因为要针对每个患者去试验哪个变异片段是有效的，这个工作量太大了，患者时间上等不及，从成本上讲也无法做到普及。

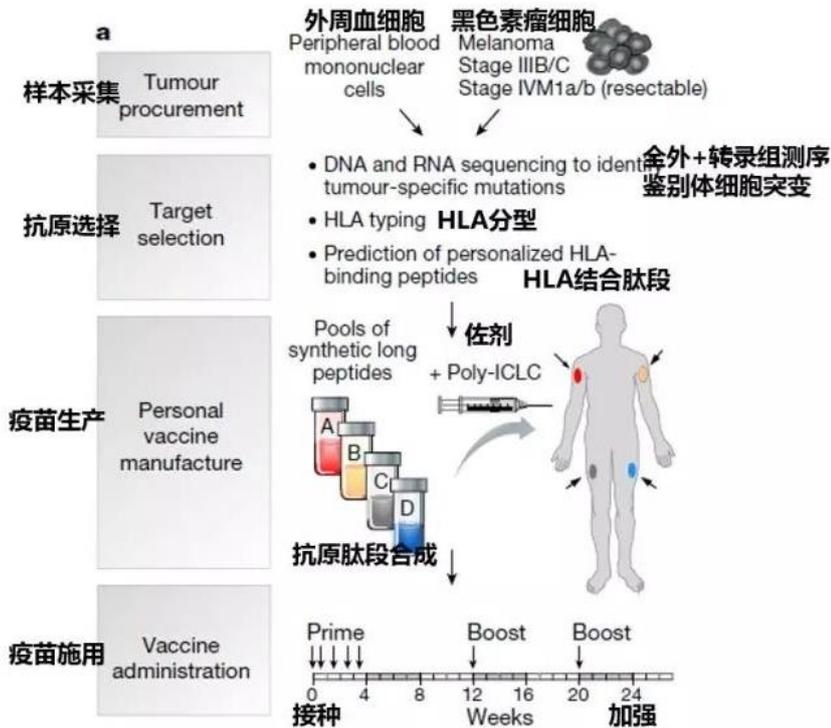


图 1. 基于个性化新抗原预测的癌症疫苗开发流程

于是，科学家们想到了人工智能，希望通过病人的基因信息来预测哪些变异片段可能作为新抗原。目前，这一方法已经取得了很大进步。如图 1 所示，通过 DNA/RNA 测序，找出癌细胞的特异性片段，再利用神经网络预测那些最有可能被免疫系统发现的片段。有了这些片段，就可以制成抗癌疫苗了【2, 3】。最近，自然-机器智能杂志也发表了一篇文章[4]，对病人肿瘤细胞表面的正常蛋白质建立深度学习模型，利用该模型判断异常蛋白质，从而发现有效的抗原。

今天，人工智能已经在医疗健康的各个领域大展身手，癌症疫苗开发只是战场之一。随着人类基因库的完善和各种医疗信息的汇集，人工智能发挥作用的场景也会越来越多，成为人类的健康的忠实守护者。

[1]Hu, Z., Ott, P. A. & Wu, C. J. Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nat. Rev. Immunol.* 18, 168–182 (2018).

[2]Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* 32, 511–517 (2016).

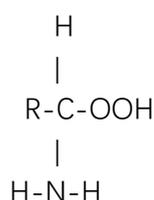
[3]Calling cancer's bluff with neoantigen vaccines, <https://www.nature.com/articles/d41586-017-08706-3>

[4] Ngoc Hieu Tran, Rui Qiao, Lei Xin, Xin Chen, Baozhen Shan, Ming Li, Personalized deep learning of individual immunopeptidomes to identify neoantigens for cancer vaccines, *Nature Machine Intelligence*, 2, pages764–771(2020)

## 相关资料：

### 什么是氨基酸，什么是蛋白质？

从有机化学讲前。有机物一般是指分子式中含有碳元素的物质。有机物的性质是由不同的基决定的。-COOH 称为羧基，是各种酸的基本功能单元。-NH<sub>2</sub> 称为氨基，是各种碱的基本功能单元。氨基和羧基通过碳元素结合在一起，成为氨基酸(amino-acid)。



取不同的 R，得到不同的氨基酸。氨基酸是生命的最小单元，是蛋白质的组成成份。

一个氨基酸的羧基和另一个氨基酸的氨基结合在一起，并去掉一分子水，得到一个肽键 (Peptide)：

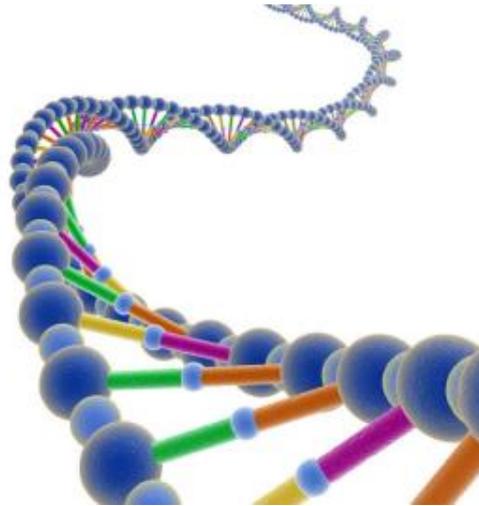


多个氨基酸通过肽键连接起来，成为一个肽链。按氨基酸的多少，肽链可以为寡肽，多肽，蛋白质 (Protein)。多肽有自己的折叠结构，结构决定功能。氨基酸的序列可决定折叠结构，进而决定其功能。AlphaFold2 已经成功实现由氨基酸序列预测蛋白质结构。

### 什么是 DNA？

碱基是碱性基团的统称，一般含有氮元素。氨基是最简单的碱基。生物体中常见的碱基有 5 种，分别是腺嘌呤 (A)、鸟嘌呤 (G)、胞嘧啶 (C)、胸腺嘧啶 (T) 和尿嘧啶 (U)。碱基通过共价键与核糖或脱氧核糖的 1 位碳原子相连而形成的化合物叫核苷。核苷再与磷酸结合就形成核苷酸，磷酸基接在五碳糖（核糖或脱氧核糖）的第 5 位 [碳原子](#) 上。

每个核苷酸单体长度为 0.33 纳米。核苷酸聚合形成长链，两条核苷酸链围绕一个共同的中心轴盘绕，由其中的碱基对互相握手，构成双螺旋结构。如果核苷酸中的糖是脱氧核糖，就得到脱氧核糖核酸，即 DNA，是生物体中最常见的遗传物质。如果核苷酸中的糖是核糖，就得到核糖核酸，即 RNA，常用于遗传物质的复制。



- DNA 中的碱基对并不都具有遗传功能。具有遗传功能的 DNA 片段称为基因 (Gene)。在进行复制时, DNA 周围会聚拢在组蛋白质周围, 这些蛋白质可被染色, 因此称为染色体 (chromosome)。因此, 染色体和基因都是功能性称呼, 并不是一种化学存在, 而 DNA/RNA/蛋白质都是。

## 人类白细胞抗原与免疫原理

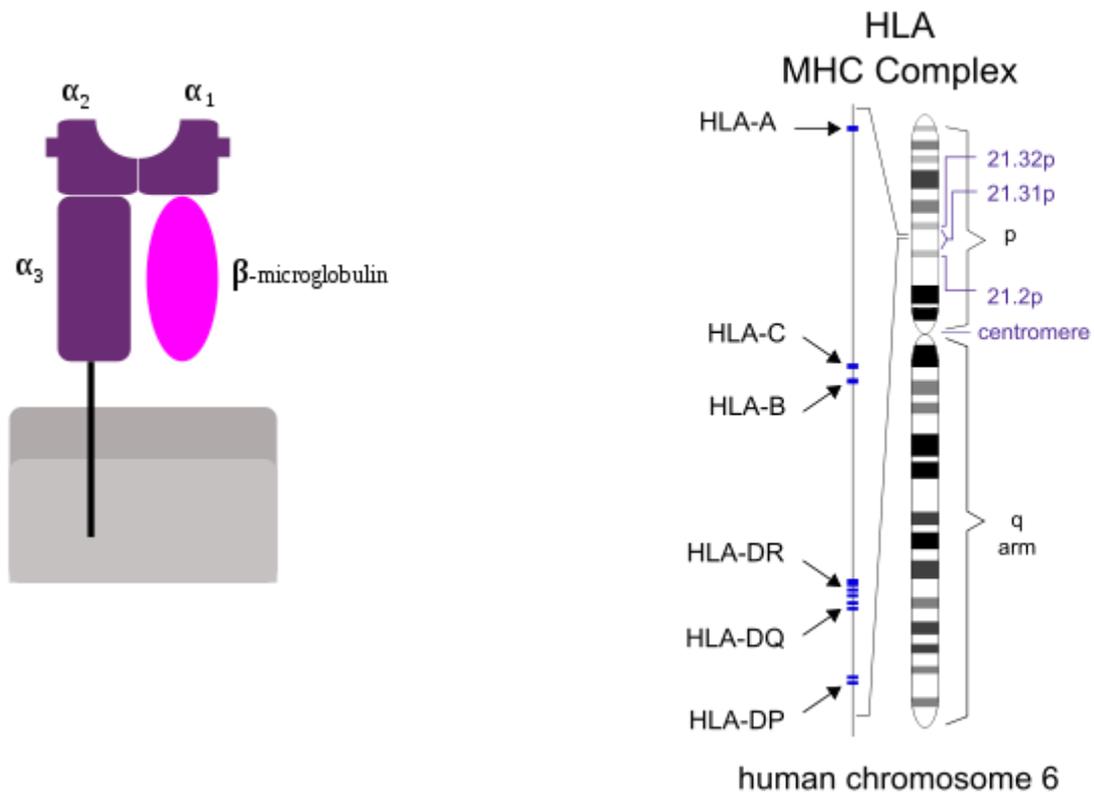
人类白细胞抗原(Human leukocyte antigen, HLA)是免疫系统的基础。

引起急性移植排斥反应的同种异型抗原称为主要组织相容性抗原(major histocompatibility antigen, MHA), 编码这组抗原的**基因**称为**主要组织相容性复合体** (major histocompatibility complex, MHC)。人类主要组织相容性抗原是在白细胞中发现的, 所以 MHA 又称为 HLA。总的来说, HLA 是一个细胞表面的蛋白质分子, 而这一分子的特性是由人类遗传物质-位于染色体中的 DNA 所编码的。

### 1.1. HLA 复合体的结构

编码 HLA 的**基因群**位于人类第 6 号染色体短臂 q21.31 和 q21.32 区之间, 分为 HLA-I 类、HLA-II 类和 HLA-III 类 3 个区, 跨度约为 4 000 kb, 约包含 450 个基因。其中经典的 HLA 基因一般专指与组织相容性有关的、具有向 T 细胞提呈蛋白质抗原功能的基因。

HLA-I类区中约包含 3 个经典的 HLA-I类基因，即 HLA-A、HLA-B、HLA-C。HLA-II类区包括 HLA-DP、HLA-DQ、HLA-DR 三个经典 HLA-II类基因及一些与抗原处理和提呈有关的基因，包括 TAP1、TAP2、LMP2 等。HLA-III类区包含 75 个基因，大多数功能不明，其中一小部分是与 HLA 无关的、编码分泌型蛋白的基因，其产物是免疫系统的重要成分。



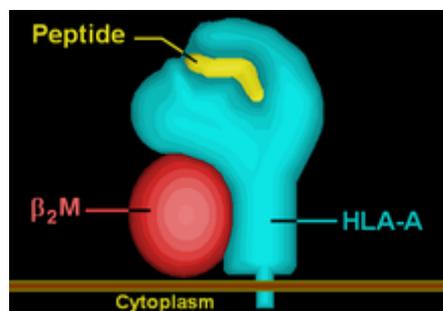
## 1.2. HLA 分子的结构和功能

HLA-I类和 HLA-II类分子是细胞表面跨膜糖蛋白，又称为 HLA-I类和 HLA-II类抗原。HLA-I类分子广泛分布于体内有核细胞膜表面，以淋巴细胞表面密度最大。与 HLA-I类分子相比，HLA-II类分子的表达更为局限，正常情况下主要表达在专职抗原提呈细胞表面，主要包括 B 细胞、单核巨噬细胞、树突状细胞。

HLA-I类分子是由一条重链和一条轻链通过非共价结合形成的异二聚体。重链又称  $\alpha$  链，由经典的 HLA-A、HLA-B、HLA-C 基因编码；轻链为  $\beta_2$  微球蛋白( $\beta_2m$ )，其编码基因位于第 15 号染色体。HLA-II类分子是由  $\alpha$  链和  $\beta$  链组成的异二聚体。 $\alpha$  链和  $\beta$  链的编码基因均位于 HLA-II类区的 D 区内。

HLA-I类分子的主要功能是作为杀伤性 CD8<sup>+</sup>T 细胞的识别标志之一，参与内源性抗原的提呈。HLA-II类分子作为 CD4<sup>+</sup>T 细胞的识别标志之一，主要参与外源性抗原的递呈。另外，一些 HLA-I类分子，包括非经典的I类分子 HLA-E 和 HLA-G，还可以与 NK 细胞表面的受体结合，影响 NK 细胞的功能。HLA 分子由于直接影响机体的细胞免疫，因而成为免疫应答过程中的关键分子。

下图是 HLA-A 分子，包含一个 alpha chain (由 HLA-A 区基因编码) 和一个 b chain，由另一区基本编码。该分子会把细胞里的碎片蛋白提到表面（黄色），免疫系统会查询这些蛋白，如果细胞被入侵，这部分蛋白会发生改变，免疫 T 细胞将会杀掉这一细胞，实现免疫。



### 1.3 免疫疗法

癌症是发生了变异的细胞。这一变异大多数情况下由 HLA 提升到细胞表面，然后被杀死。如果变异无法被提升到表面，或提升到表面但免疫系统无法发现，则变异细胞会存活下来，成为癌细胞。免疫疗法的原理在于定位那些在 HLA 结合的蛋白质中变异的片段，从而激发免疫系统。

CAR-T 细胞疗法是将 T 细胞从病人体内提取出来进行基因编辑，让它识别肿瘤特征，再送回病人体内。这一方法的关键在如何让 T 细胞识别肿瘤特性，这是很难的事。

目前，CAR-T 疗法在临床上取得的成功，大都是在由 B 细胞发展成的肿瘤中，比如 B 细胞非霍奇金淋巴瘤（B-cell non-Hodgkin lymphoma, NHL）、弥漫性大 B 细胞淋巴瘤（Diffuse large B-cell lymphoma, DLBL）以及急性 B 淋巴细胞白血病（B-ALL）。这是因为被现有技术改造的 T 细胞，能特异识别 B 细胞表面抗原，从而把病人体内所有的 B 细胞——不管是癌还是正常细胞——一股脑儿全都消灭掉。毕竟，一个人即使没有了正常的 B 细胞还是可以活下去的，这么做问题不大。

作者：返朴

链接：<https://www.zhihu.com/question/26897722/answer/718529946>

来源：知乎

著作权归作者所有。商业转载请联系作者获得授权，非商业转载请注明出处。

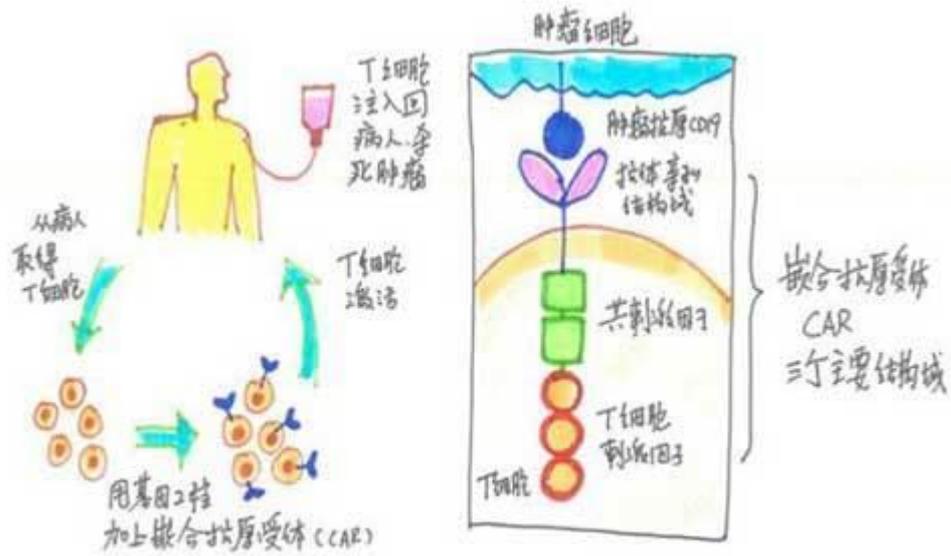


图3 嵌合抗原受体(CAR)自体T细胞疗法

插图改编自《自然》杂志, 2013年504卷

杨青 2014.4. 画