

网络平台上的大规模化合物结构检索解决方案

中国科学院过程工程研究所多相复杂系统国家重点实验室

高性能计算与化学信息学课题组

网站: <http://chemport.ipe.ac.cn/IPE-ChINGroup/index.html>

概况: 化学结构是专业人员直观、唯一地表示化合物的手段,近年来,作为化学信息学的一类核心技术,国际上化学结构检索技术逐步趋于成熟,广泛应用于计算机辅助新药开发、化学数据库、企业化学品目录的检索。由于我国化学结构检索技术研发的严重滞后,到目前为止我国在药物研发方面的化学结构信息处理主要依赖国外软件;化学结构检索在化学数据库、企业化学品目录方面的应用基本上是空白。我所高性能计算与化学信息学课题组在建立网络化学化工信息资源集成检索平台系列工具的过程中,建立了一套面向网络化学数据库和企业网站化学品目录的结构检索系统,为实现网络化的化合物的全结构和子结构的检索提供了一套完整的解决方案。

技术特点: 我们的化合物的全结构和子结构的检索利用国际著名的化学软件工具包 CDK、JOELib 的基本构件,基于 Java、并综合运用多种先进的网络化信息技术(JSP+ Servlet+Bean+MySQL)实现。检索输入可以是在线绘制的 2D 分子结构图、或在线提交的分子结构 Mol 文件;检索结果实现了可操控的 3D 分子结构模型显示。由于化合物子结构匹配是一个 2D 子图匹配问题,匹配时间可随着原子规模的扩大呈指数增长,我们采用了根据化合物的结构特征和指纹图谱(fingerprint)进行预筛选、然后进行严格匹配的策略,有效解决了化学子结构检索的效率问题。结合多线程优化和小型集群并行,应用于规模为万级、十万级、到百万级的化学结构检索均有令人满意的效果。

实际应用情况: 开发和内部测试已经完成,已经应用到我所的化学深层网检索引擎 ChemDB Portal(<http://www.chemdb-portal.cn/>)中。

市场分析: 国际上发展的趋势是化学结构越来越多地应用于网络化学数据库的检索、企业网站中产品目录的检索,目前国内应用几乎是空白。我们的化学结构及子结构检索技术为国内网络化学数据库和企业网站中产品目录的检索方式向结构检索的升级、与国际同类应用接轨提供了一个经济、方便的解决方案。

合作方式: 我所提供完整的技术解决及实施方案。