竞赛介绍

一、介绍

全球新冠肺炎确诊病例逾千万，且仍以每天30万新增病例飙升，一些发展中国家正在成为新“震中”，形势堪忧。新冠病毒毒性大，隐密性、变异性强，扩散快，稍有松懈便可酿成大患，后疫情时代，新型冠状病毒基因组演化分析，为全面评估疫情风险、启动公共卫生应对措施及制定医疗对策提供全面、有效的数据支撑。
隐私计算，广义上是指带有隐私信息保护的计算系统与技术（包括但不限于联邦学习（Federated Learning, FL）、可信计算环境（Trusted Execution Environment, TEE）、安全多方计算（Secure Multiparty Computation, MPC）、同态加密（Homomorphic Encryption, HE）等技术），其能够在不泄露原始数据的前提下，对数据进行分析处理验证，包括数据的生产、存储、计算、应用等信息处理流程的全过程。
使用隐私计算技术可全面打通各数据源之间的数据孤岛，突破数据分析门槛，提升数据利用率，促进多中心合作共享及成果转化。

二、赛题任务

- 计算参与方：甲方、乙方共两方。
- 数据输入：
  - 长度为K的基因序列集，甲方有M条基因序列，乙方有N条基因序列。
    - 测评时基因序列的长度K的取值分三个档次，分别在0.5k，5k 和30k数量级。
    - 测评时M的取值范围：500条 ~ 30000条
    - 测评时N的取值范围：500条 ~ 70000条
  - 大赛考虑两种数据预处理情况：
    - PP1：甲乙双方基因序列对输入前已经完成对齐（alignment）
      - 大赛主办方预先采用成熟的软件mafft进行比对，确保序列对齐的准确性，然后进行数据截取。
        https://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html
    - PP2：甲乙双方基因序列对输入前未完成对齐，需要通过隐私计算完成甲乙双方基因序列对对齐（alignment）。
      - 具体算法请参考：“具体算法->甲乙双方多重基因序列对齐”部分。
- 目标输出：根据指定的基因序列对之间距离计算算法，生成一个基于甲乙

双方基因序列对之间的距离矩阵，并根据该距离矩阵，使用NJ算法（neighbor joining），计算出一颗结果NJ树，计算结果仅由甲方获得。

- 技术要求 :基于如下三种技术中的一种构建可实测的多方协同隐私计算软件程序，禁止在一个程序中混合使用多种隐私计算技术。
  - T1：同态加密（HE）
  - T2：安全多方计算（MPC）
  - T3：基于SGX V1的可信计算环境（TEE）
- 安全性要求：128bits安全性。
  - 对于T1:同态加密（HE）解决方案：参赛队伍请参考国际同态加密标准白皮书5.4 TABLES of RECOMMENDED PARAMETERS中推荐的参数设置。http://homomorphicencryption.org/white_papers/security_homomorphic_encryption_white_paper.pdf
  - 对于T2: 多方安全计算：可以使用乱码电路或机密分享算法，如使用机密分享算法，需要明确声明(t, n)阈值密钥分享的设定。https://en.wikipedia.org/wiki/Secret_sharing
  - 对于T3: 可信计算环境(TEE)：参赛队伍请参考SGX Remote Attestation协议完成动态密钥交换。本赛题暂不考虑测信道攻击问题。https://www.intel.com/content/www/us/en/developer/articles/code-sample/software-guard-extensions-remote-attestation-end-to-end-example.html
- 安全假设：
  - 多方安全计算（MPC）和同态加密（HE）解决方案需要支持半诚实模型假设；
  - 基于可信计算环境（TEE）的解决方案需要支持恶意模型假设；
- 隐私保护目标：
  - **PG1:** 任一参与方的基因序列集数据明文不能被其他参与方以截获的方式获知。但是甲乙双方的序列之间的相似矩阵可以作为结果的一部分输出，结果NJ树可以保留每条边的权重信息。
  - **PG2**：任一参与方的基因序列集数据明文不能被其他参与方以截获的方式获知，结果NJ树仅需保留节点和边的树形结构（每条边上权重明文不输出，避免通过权重反推出距离矩阵和相似矩阵）
- 评测原则：
  - 不同的隐私计算技术解决方案（T1：HE，T2：MPC和T3：TEE）将分开测评；

    - 不同的隐私保护目标（PG1和PG2）将分开测评；
    - 不同的数据预处理假设（PP1和PP2）将分开测评；
    - 对于基于T1：HE 或 T2：MPC的解决方案
        - 测评时每个参赛队伍的解决方案必须要指定唯一的一种隐私保护目的（也就是PG1或PG2）和唯一的一种数据预处理假设（也就是PP1或PP2）。
    - 对于T3：TEE解决方案
        - 解决方案只能采用PG2隐私保护目的和PP2数据预处理假设。
    - 测评时允许每个参赛队伍基于不同的隐私计算技术（T1，T2和T3）提交多个隐私计算软件程序，但是每个软件程序不能混用HE、MPC或TEE技术。
    - 大赛主办方将基于保留的未公开数据进行测试。
- 评测指标：
    - 结果的正确性（基于隐私计算和明文计算下的NJ树节点和边的树形结构对比）。
    - 联合计算所需的总时间（包括代码编译、预处理、联合计算等）。
    - 计算过程中的内存使用量峰值。
    - 计算过程中的网络通信总流量。
    - 计算程序所能支持的基因序列长度K和所能支持的甲乙双方的基因序列条数M和N。
- 评测环境 :提供相对高配的计算型云主机配置，通过Docker部署，Docker计算环境的参考配置为：4核心CPU，32 GB 内存，500GB硬盘。
- 具体算法:
    - 甲乙双方多重基因序列对齐:
        - 通过Needleman and Wunsch algorithm 算法，匹配为2分，不匹配-1分，gap为-3分，回溯寻找两两最优匹配结果方案。详细可参考 https://en.wikipedia.org/wiki/Needleman%E2%80%93Wunsch_algorithm
        - NW算法只能进行两两比对，多序列比对是采用渐进式，逐条与其他序列比对，直到构建对齐序列文件。
        - 具体算法请参考：code-COVID-evolution.html（1.多重序列比对）
    - 构建相似性矩阵：
        - 序列相似性矩阵评估有多种模型，示例参考 p-distance.更多详细请看

https://www.megasoftware.net/mega1_manual/Distance.html

        - 具体算法请参考：code-COVID-evolution.html （2.构建相似性矩阵）
    - 构建进化树：
        - 采用NJ方法进行构树
        - 具体算法请参考：code-COVID-evolution.html（3.构建进化树）
- 示例数据：大赛方提供如下公开数据供参赛队伍开发解决方案。
    - 已经对齐的数据：
        - L1: K = 500; M+N = 1000;
        - L2: K = 5000; M+N = 2000;
        - L3: K = 20000; M+N = 20000;
    - 未对齐的数据：
        - L1: K = 0.5k数量级； M+N = 1000;
        - L2: K = 5k数量级； M+N = 2000;
        - L3: K = 30k数量级; M+N = 20000;

	PP1:序列已对齐	PP2：序列未对齐
PG1：保护原序列	HE/MPC	HE/MPC
PG2：保护原序列和相似矩阵	HE/MPC	MPC/TEE