ITEM #236 - Two-Ways CCC Test Algorithm for Functional DNA Signal Detection

Conversation：Two-Ways CCC Test

20260114

Authors: Sizhe Tan & GPT-Obot

DBM-COT ITEM #236 (EN)

Two-Ways CCC Test Algorithm for Functional DNA Signal Detection

1. Motivation and Problem Statement

Mapping **DNA sequences** to **biological phenotypes or functions** (e.g. facial traits, morphology, disease risks) is fundamentally a **weak-signal discovery problem** under heavy noise, combinatorial effects, and ambiguous labels.

Traditional regression or end-to-end learning approaches tend to:

- entangle background genomic structures with functional signals;
- overfit to spurious correlations;
- provide limited interpretability and weak error diagnostics.

This ITEM proposes a **Two-Ways CCC Test Algorithm**, a **structure-first, bidirectional validation framework**, designed to:

- isolate functional DNA expression patterns;
- explicitly separate *positive*, *negative*, and *background* signals;
- quantify confidence via symmetric false-negative / false-positive analysis;
- support iterative refinement of phenotype definitions.

---

2. Input Observations

We assume a large observational dataset:

O={ (Si, Ai)∣i=1,…,N }\mathcal{O} = \{\, (S_i,\ A_i) \mid i=1,\dots,N \,\}O={(Si, Ai)∣i=1,…,N}

where:

- SiS_iSi is a DNA sequence;
- Ai={Attributei,j}A_i = \{Attribute_{i,j}\}Ai={Attributei,j} is a set of observed biological attributes.

Given a **target functional attribute subset**:

T⊆Attribute spaceT \subseteq \text{Attribute space}T⊆Attribute space

the task is to discover **DNA expression patterns (CCC structures)** that reliably encode, promote, or suppress TTT.

3. Three-Group Partitioning

All observations are partitioned into three disjoint groups:

- **Group 1 (Positive group)**

  G1={i∣T⊆Ai}G_1 = \{ i \mid T \subseteq A_i \}G1={i∣T⊆Ai}

  Observations fully expressing the target function.

- **Group 2 (Negative group)**

  G2={i∣Ai∩T=∅}G_2 = \{ i \mid A_i \cap T = \emptyset \}G2={i∣Ai∩T=∅}

  Observations completely lacking the target function.

- **Group 3 (Partial / ambiguous group)**

  G3=all remaining observationsG_3 = \text{all remaining observations}G3=all remaining observations

  Observations partially matching the target attributes.

4. Core Algorithm (Two-Ways CCC Test)

Step 1: Independent CCC Extraction

- Compute **CCC_1** from DNA sequences in G1G_1G1.
- Compute **CCC_2** from DNA sequences in G2G_2G2.

Each CCC represents a set of recurrent, structured DNA expression patterns.

Step 2: Background Elimination via Bidirectional Intersection

- Compute:

  IntersectCCC=UnalignedAND(CCC1, CCC2)\text{IntersectCCC} = \text{UnalignedAND}(CCC_1,\ CCC_2)IntersectCCC=UnalignedAND(CCC1, CCC2)

These structures appear frequently **regardless of function presence** and are treated as **background or generic genomic patterns**.

Step 3: Signal Separation

- **Positive functional signals**

  CCC+=CCC1\IntersectCCCCCC^+ = CCC_1 \setminus \text{IntersectCCC}CCC+=CCC1\IntersectCCC

- **Negative functional signals**

  CCC−=CCC2\IntersectCCCCCC^- = CCC_2 \setminus \text{IntersectCCC}CCC−=CCC2\IntersectCCC

This yields **directional evidence**:

- CCC^+ → structures supporting the function;
- CCC^- → structures suppressing or excluding the function.

---

5. Bidirectional Error Estimation

Define a detection predicate:

hit(Si,X)∈{0,1}\text{hit}(S_i, X) \in \{0,1\}hit(Si,X)∈{0,1}

indicating whether DNA sequence SiS_iSi expresses CCC signal set XXX.

5.1 Missing-Positive Error (False Negative)

Using CCC$^+$:

FNR=∑i∈G1(1−hit(Si,CCC+))∣G1∣\text{FNR} = \frac{ \sum_{i \in G_1} (1 - \text{hit}(S_i, CCC^+)) }{ |G_1| }FNR=∣G1∣∑i∈G1(1−hit(Si,CCC+))

This estimates how often the functional signal is **missed** when the function is present.

5.2 False-Positive Error

FPR=∑i∈G2hit(Si,CCC+)∣G2∣\text{FPR} = \frac{ \sum_{i \in G_2} \text{hit}(S_i, CCC^+) }{ |G_2| }FPR=∣G2∣∑i∈G2hit(Si,CCC+)

This estimates how often the signal appears **without the function**.

Symmetric statistics can be computed for CCC$^-$ if needed.

---

6. Confidence Scoring and Ranking

A recommended composite score:

Score=(TPR−FPR)⋅log(1+TP+FP)\text{Score} = (\text{TPR} - \text{FPR}) \cdot \log(1 + TP + FP)Score=(TPR−FPR)⋅log(1+TP+FP)

where:

- TPR = 1 − FNR;
- TP, FP are hit counts.

This balances:

- discriminative power;
- signal coverage;
- robustness against small-sample artifacts.

Top-K CCC signals are selected based on this score.

---

7. Role of Group 3 (Partial Observations)

Group 3 provides **structural continuity information** rather than binary labels.

Recommended uses:

- **Soft weighting** based on partial attribute match ratios;
- **Conflict detection**:
    - high attribute match but no CCC$^+$ hit → possible phenotype noise;
    - low attribute match but strong CCC$^+$ hit → hidden sub-phenotype.

These conflicts drive **attribute refinement and sub-phenotype discovery**.

---

8. Attribute Refinement Loop

The algorithm supports iterative improvement of phenotype definitions:

- Attributes whose removal improves Score are flagged as noisy;
- Conflict-heavy samples are clustered to propose:
    - refined attributes;
    - composite or conditional attributes;
    - alternative biological interpretations.

Thus, **attribute definitions co-evolve with discovered DNA structures**.

---

9. Key Properties

- **Bidirectional validation**: positive and negative evidence are both required.
- **Background-robust**: generic genomic patterns are explicitly eliminated.
- **Explainable**: CCC structures form interpretable evidence chains.
- **Iterative**: naturally supports scientific hypothesis refinement.

---

10. Position in DBM-COT

This ITEM establishes a **foundational functional-signal discovery primitive** for:

- DNA → phenotype mapping;
- genotype-function hypothesis ranking;
- integration with CCC-based rule engines and hypothesis exporters.

DBM-COT ITEM #236（中文）

用于生物功能 DNA 信号发现的双向 CCC 验证算法

1. 背景与问题定义

DNA 序列到生物表型（如人脸特征、形态性状、疾病风险）的映射，本质上是一个**弱信号** + **高噪声** + **多因果组合**的问题。

传统回归或端到端 AI 方法往往：

- 将功能信号与背景基因结构混杂；
- 对偶然相关过拟合；
- 缺乏清晰的错误诊断与结构解释能力。

本 ITEM 提出 **Two-Ways CCC Test Algorithm（双向 CCC 验证算法）**，其目标是：

- 从结构层面发现 DNA 的功能表达模式；
- 显式区分正信号、负信号与背景结构；
- 通过漏报率 / 错报率给出可比较的置信度；
- 支持表型定义的迭代修正与细分。

---

2. 输入观测模型

给定观测集合：

O={(Si, Ai)|i=1,…,N}\mathcal{O} = \{ (S_i,\ A_i) \mid i=1,\dots,N \}O={(Si, Ai)|i=1,…,N}

其中：

- SiS_iSi：DNA 序列；

- AiA_iAi：该样本的生物属性集合。

选定目标生物功能属性子集：

T⊆AttributeT \subseteq AttributeT⊆Attribute

目标是寻找**与 TTT 强相关的 DNA 结构化表达模式（CCC）**。

---

3. 三组划分（Three-Group Split）

- **Group 1（功能存在）**
  完全满足 TTT 的样本；
- **Group 2（功能缺失）**
  与 TTT 完全无交集的样本；
- **Group 3（部分满足）**
  其余部分匹配的样本。

这是算法后续“对称验证”的基础。

---

4. 双向 CCC 核心流程

4.1 CCC 提取

- 在 Group 1 上生成 CCC$_1$；
- 在 Group 2 上生成 CCC$_2$。

4.2 背景结构剔除

计算：

IntersectCCC=CCC1∩CCC2\text{IntersectCCC} = CCC_1 \cap CCC_2IntersectCCC=CCC1∩CCC2

这些结构在“有功能 / 无功能”中都高频出现，视为**非功能性背景结构**。

4.3 信号分离

- **正向功能信号**

  CCC+=CCC1∖IntersectCCCCCC^+ = CCC_1 \setminus \text{IntersectCCC}CCC+=CCC1∖IntersectCCC

- **反向功能信号**

  CCC−=CCC2∖IntersectCCCCCC^- = CCC_2 \setminus \text{IntersectCCC}CCC−=CCC2∖IntersectCCC

---

5. 漏报率与错报率估计

定义命中判定：

hit(Si,X)\text{hit}(S_i, X)hit(Si,X)

漏报率（Missing-Positive）

FNR=#{i∈G1∣未命中 CCC+}∣G1∣\text{FNR} = \frac{\#\{ i\in G_1 \mid \text{未命中 } CCC^+ \}}{|G_1|}FNR=∣G1∣#{i∈G1∣未命中 CCC+}

错报率（False-Positive）

FPR=#{i∈G2∣命中 CCC+}∣G2∣\text{FPR} = \frac{\#\{ i\in G_2 \mid \text{命中 } CCC^+ \}}{|G_2|}FPR=∣G2∣#{i∈G2∣命中 CCC+}

---

6. 置信度评分与 Top-K 选择

推荐综合评分：

Score=(TPR−FPR)·log⁡(1+TP+FP)\text{Score} = (\text{TPR} - \text{FPR}) \cdot \log(1 + TP + FP)Score=(TPR−FPR)·log(1+TP+FP)

用于对不同 DNA 功能信号候选进行排序。

---

7. Group 3 的作用

Group 3 不直接参与二分类，而用于：

- 连续性验证；
- 冲突样本识别；
- 表型噪声与隐藏子功能的发现。

---

8. 属性定义的自我修正

通过：

- 去除导致 Score 下降的属性；
- 分析 Group 3 冲突子群；

可以自动发现：

- 坏的表型定义；
- 需要细分的新属性；
- 条件型或组合型生物功能。

---

9. 算法特性总结

- **双向验证**：必须同时通过正向与反向检验；
- **抗背景噪声**：显式剔除共性结构；
- **可解释**：CCC 作为结构化证据链；
- **可演化**：属性定义与结构共同迭代。

---

10. 在 DBM-COT 中的位置

本 ITEM 为 **DNA 功能信号发现与假设生成** 提供了一个核心、稳定、可复用的结构智能原语。
