Aquatic Bacteria Database

Or:

Using an established knowledge base to assist the Institute of Aquaculture's diagnostic services

The Institute & Diagnosis

- IoA maintains large database of bacteriological bio-chemical test results.
- Academics offer confidential consultancy worldwide to identify potential/specific bacteria given an organisation's test results.
- Correct ID to inform health & safety concerns, e.g. disease-causing bacteria, hence diagnosis.
- Presently diagnosis is conducted manually(!).

Test Results to Diagnosis: Examples

Knowledge Base

Strain	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7
A	+	+	+	-	-	-	-
B	+	+	-	+	+	-	-
C	+	-	-	+	-	+	-

Clients' Samples

Sample	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7
C1	+	+	+	?	?	?	?
C2	+	+	?	?	?	-	-
C3	?	+	-	-	+	?	?

Samples supplied are incomplete: Clients constrained by their own resources. We evaluate each sample in turn...

Exact Match: Strain A

Knowledge Base

Strain	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7
A	+	+	+	-	-	-	-
B	+	+	-	+	+	-	-
C	+	-	-	+	-	+	-

A precise, singular match!

Clients' Samples

Sample	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7
C1	+	+	+	?	?	?	?
C2	+	+	?	?	?	-	-
C3	?	+	-	-	+	?	?

Multiple Matches: Strains A or B

Knowledge Base

Strain	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7
A	+	+	+	-	-	-	-
B	+	+	-	+	+	-	-
C	+	-	-	+	-	+	-

Supplied samples eliminate w/in KB, but not to singular results.

Clients' Samples

Sample	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7
C1	+	+	+	?	?	?	?
C2	+	+	?	?	?	-	-
C3	?	+	-	-	+	?	?

No Matches

Knowledge Base

Strain	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7
A	+	+	+	-	-	-	-
B	+	+	-	+	+	-	-
C	+	-	-	+	-	+	-

?

Though some tests correspond, others do not, clearly no exact match!

Clients' Samples

Sample	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7
C1	+	+	+	?	?	?	?
C2	+	+	?	?	?	-	-
C3	?	+	-	-	+	?	?

Implementation

- Functions to include:
  I. Search/diagnose
  II. Maintain/add to KB
  III. Useable GUI (likely via Java/Python)
  IV. Admin & Inquirer user types
  V. Advise further tests to narrow to singular result
  VI. Vary ‘threshold of accuracy’ of diagnosis

Goals & Extension

- Measures of Success:
  1. It produces correct diagnosis
  2. Product functionality (of *I.-VI.* noted).
- Extension:
  - Web pages w/PHP script to interface to DB/KB.
  - Case-based learning approach.

Considerations on Diagnosis

- Diagnosis only as good as tests:

MORE = MERRIER

- Difficulties in progress:
  - Coherency of KB: Is it accurate to the biology?
  - Advising: not all ‘no matches’ are ‘new strains’ in no-match case.
  - Capturing expert-like insight in complex cases.
    - How to recommend further tests , how to integrate cost/efficacy.
    - Embedding significance of certain tests.
      - It is known ~6 particular tests are usually enough to ID Genus alone.
  - Effectively Boolean test results:
    - Revisions to bio-chemical test specifications could effect/invalidate KB.
    - Ideal would be to take raw *numeric* values and permit tuning of test conditions in the product.

ANY QUESTIONS?

Image: http://wearscience.com

Appendix A: Diagnosis Pseudocode

Client tests input I[t] to be checked against knowledge base KB[r,t]
Create M which will hold matched results

```
ForAll t {
IF isMatch(I, KB[r]) THEN
        add KB[r] to M
} return M
```

Def^{N}: isMatch(I, KB[r]) = {
True: forall t. IF ((I[t] == KB[r,t]) OR (I[t] = '?'))
False: Otherwise
}

```
IF (M is empty)
   @user:   "No matches found for your test results"
ELSE IF (M is only one entry)
   @user:   "Unique match found. Here are the details..."
ELSE IF (M is many)
   @user:   "Multiple possible matches found."
   FOR each entry in M
      @user:   "<details of each entry>"
```

t:	Test
r:	'Row': genus-species-strain
?:	Not supplied
KB[r]:	Strain
KB[r,t]:	Test result of t, strain r
I[t]:	Sample test result
M:	Holds all KB(r) which match

Appendix B: Assumptions

i. **The KB aspires to completeness: all entries have full, unambiguous test results. All new entries must obey this.**
   i. This can be altered. It is possible to accommodate incomplete entries, to work with partial information. Awaiting confirmation from Dr Crumlish on the generality of such assumptions

ii. **Strains are coherent: a strain's profile cannot be duplicated by another distinct strain.**

iii. **Strains have one profile: the same strain cannot yield two distinct profiles.**

iv. **Diagnosis is conducted one sample at a time.**
   i. Again, it is possible to imagine methods of input (e.g. via correctly prepared standardised spreadsheets) which can be accessed easily allowing users to diagnose more than one sample at one time. (Harness the power of computing!)

v. **A match is a valid if all known and specified tests of the input tests are correctly identified to strain(s) in the KB.**
   i. We can conceive of this being scalable: a user may elect for, say, 'only 80% of inputs' to qualify for the diagnosis, given some reason for the uncertainty. Moreover, output could be a scale of 'most accurate matches' down to a lower threshold (even as far as simply ordering the KB with 'most accurate matches' given first.) Even more, it could be imagine a client could ask "Can we be sure >sample< is not >subset of strains<?" and it would be worthwhile to say "This sample does not seem to be >subset of strains<".