Protein identification using mass spectrometry data and the case of the glass slipper

Malcolm Farrow and Gavin Hope
University of Sunderland
UK

Nonlinear Dynamics Ltd.
UK

Key reference

Grimm, J.L.C. and Grimm, W.C., 1857, *Kinder- und Hausmärchen* (Children's and Household Tales), 7th Edition.

1. Protein identification from a Bayesian viewpoint.

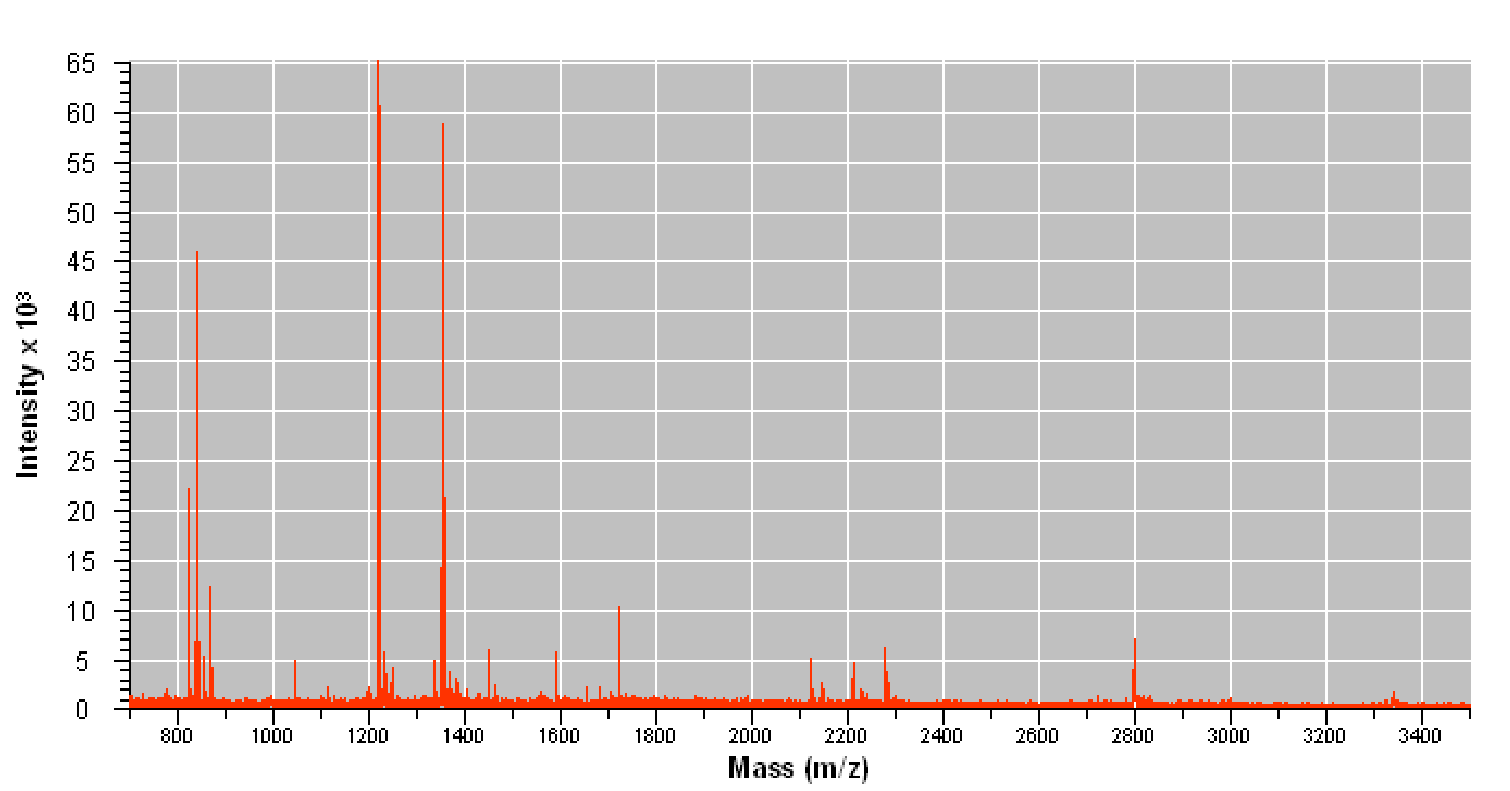
2. Complications and difficulties.

3. Review of some available algorithms.

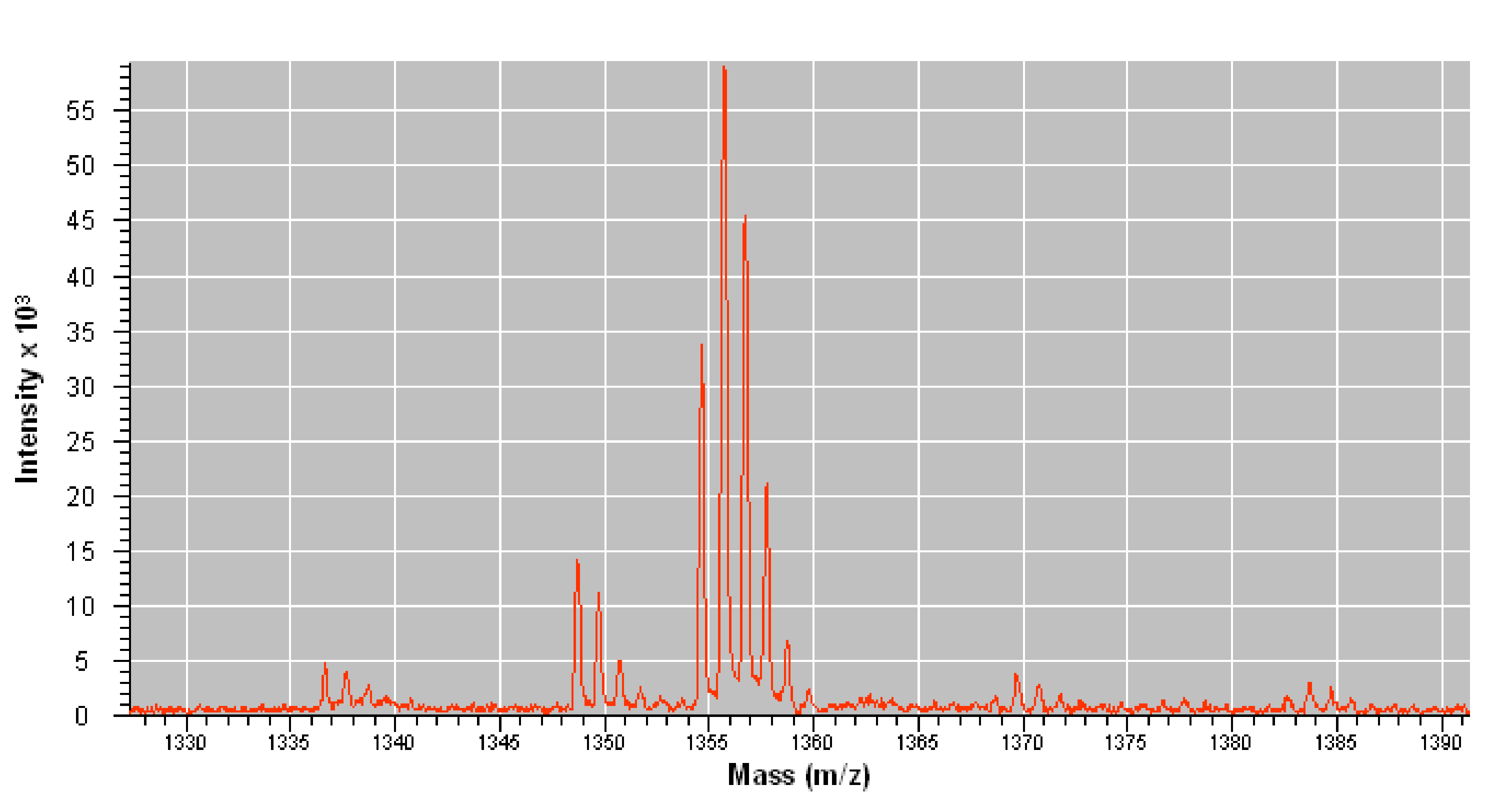
4. Dealing with the difficulties.

5. The “glass (or golden!) slipper” problem.

Picture of a spectrum.



Close-up of a spectrum.



General Principle

- Sample of some unidentified protein k^*.
- Database $S = \{k_1, \ldots, k_n\}$ of known proteins.
- Prior probability that $k^* = k_i$ is $p_i^{(0)}$.

$$\Pr(k^* = k_i | y) = p_i^{(1)} = \frac{p_i^{(0)} L_{yi}}{\sum_{j=1}^{n} p_j^{(0)} L_{yj}}. \qquad (1)$$

Application to the protein problem

- For the sample we have a set P_y of n_y observed peaks.
- For each k_i we have a set P_i of n_i theoretical peaks.
- Let $n_{yi} = \min\{n_y, n_i\}$. The number of possible allocations is

$$N_{yi} = \sum_{m=0}^{n_{yi}} \frac{n_y! n_i!}{m!(n_y - m)!(n_i - m)!}.$$

- The likelihood is

$$L_{yi} = \sum_{j=1}^{N_{yi}} \Pr(a_{yij}|k_i = k^*) \Pr(P_y|k_i = k^*, a_{yij}). \qquad (2)$$

Complications and difficulties

- “Noise”.
- Location shifts of peaks.
- Theoretical peaks might not appear. (Missed cleavage, obscured by noise, not ionised etc.)
- Unexpected peaks might appear. (E.g. contamination, modifications).

Some algorithms

- Probabilistic
- Other

“Probabilistic”

- “MOWSE” (Pappin, Højrup and Bleasby, 1993)
- “Mascot” (Perkins, Pappin, Creasy and Cottrell, 1999)
- “MSROFIT” (Berndt, Hobohm and Langen, 1999)
- “ProFound” (Zhang and Chait, 2000)

Other

- “PepSea” (Mann, Højrup and Roepstorff, 1993)
- “PeptIdent” / “MultiIdent” (Wilkins *et al.*, 1998, 1999)
- “PeptIdent2” (Gras *et al.*, 1999)

See also, e.g., Fenyö (2000)

What probabilities?

- *Cf.* forensic DNA database search problem. E.g. Balding (2002).

- - Posterior $\mathrm{Pr}(k^* = k_i \mid y)$.
  - Probability, given y, that we will find a "match" in the database "by chance."
    - "significance"
    - "false positive probability"

- $Pr(y \mid k^* = k_i)$ or $\mathrm{Pr}($ “match” $\mid y)$.

- $\mathrm{Pr}(k^* \notin S)$. See later.

Plug-in probabilities

Example: Given θ,

$$p = \Pr(\text{peak appears} \mid \theta) = \theta.$$

$$\theta \sim \text{beta}(a, b)$$

One peak:

$$p = \int_0^1 \theta f(\theta)\ d\theta = \frac{a}{a+b}.$$

Two peaks:

$$\int_0^1 \theta^2 f(\theta)\ d\theta = \left(\frac{a}{a+b}\right)\left(\frac{a+1}{a+b+1}\right) > p^2.$$

Sequence of x appearances and $n-x$ non-appearances:

$$\begin{aligned}
p_n(x) &= \int_0^1 \theta^x (1-\theta)^{n-x} f(\theta)\ d\theta \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+x)\Gamma(b+n-x)}{\Gamma(a+b+n)} \\
&= \frac{C(a,x)C(b,n-x)}{C(a+b,n)}
\end{aligned}$$

$$p_n(x) = \frac{C(a,x)C(b,n-x)}{C(a+b,n)} \tag{3}$$

where

$$\begin{aligned} C(a,x) &= a(a+1)\cdots(a+x-1) \\ C(b,n-x) &= b(b+1)\cdots(b+n-x-1) \\ C(a+b,n) &= (a+b)(a+b+1)\cdots(a+b+n-1) \end{aligned}$$

Dealing with the difficulties

- Peak extraction
- Appearance of predicted peaks
- Location shifts of observed peaks
- Appearance of extra peaks

$$\begin{aligned}\Pr(a_{yij} \mid k_i = k^*) &= \Pr[b(a_{yij}) \mid k = k^*] \\ &\quad \times \Pr[c_y(a_{yij}) \mid k_i = k^*]\end{aligned}$$

$b(a_{yij})$: exactly the selection of species from P_i required by the allocation a_{yij} appears. See (3).

$\Pr[c_y(a_{yij}) \mid k_i = k^*]$: the probability density for the observed peaks appearing in their observed locations, given the allocation a_{yij}.

Location shifts

D_m

Model 1.

$$\begin{aligned}
\mathrm{E}(D_m) &= 0 \\
\mathrm{var}(D_m) &= \sigma_c^2 + \sigma_e^2 \\
\mathrm{covar}(D_m, D_{m'}) &= \sigma_c^2
\end{aligned}$$

Model 2.

Internal calibration at two masses, c_1, c_2.

Two theoretical masses, m, m'.

$$c_1 < m,\ m' < c_2$$

Adjustments A_1, A_2 made at masses c_1, c_2.

Adjustment made at m :

$$\frac{A_1(c_2 - m) + A_2(m - c_1)}{c_2 - c_1}$$

but A_1, A_2 have error – assume independent here.

D_m is adjustment error plus error specific to m.

$$
\begin{aligned}
\mathrm{var}(D_m) &= C(m,m)\sigma_c^2 + \sigma_e^2 \\
\mathrm{covar}(D_m, D_{m'}) &= C(m,m')\sigma_c^2
\end{aligned}
$$

$$
\begin{aligned}
C(m,m) &= \frac{(c_2 - m)^2 + (m - c_1)^2}{(c_2 - c_1)^2} \\
C(m,m') &= \frac{(c_2 - m)(c_2 - m') + (m - c_1)(m' - c_1)}{(c_2 - c_1)^2}
\end{aligned}
$$

Models 3, 4

(May have to transform masses).

Theoretical mass	Observed mass	Calibration masses
t_i	z_i	$c_1 < \cdots < c_s$

Also $c_0 \equiv 0$ (usually).

Theoretical masses may be

$$c_j < t_i < c_{j+1} \qquad c_s < t_i$$

Model 3

Before calibration:

$$z_i \sim N(t_i,\ t_i\sigma_c^2 + \sigma_e^2)$$

$$\mathrm{covar}(z_i, z_j) = \sigma_c^2 \min(t_i, t_j)$$

Similarly obs. values for calibration masses.

Condition on observations of calibration masses.

Model 4.

$$z_i \sim \mathrm{gamma}(t_i\lambda, \lambda) \quad (\lambda > 0)$$

$$\mathrm{E}(z_i) = t_i$$

$$z_2 - z_1 \sim \mathrm{gamma}([t_2 - t_1]\lambda, \lambda) \quad (t_1 < t_2)$$

Joint density of obs. masses (inc. calib. masses).

$$\prod_{i=1}^{n} \left\{ \frac{\lambda^{w_i}(z_i - z_{i-1})^{w_i-1} e^{-\lambda(z_i - z_{i-1})}}{\Gamma(w_i)} \right\} \tag{4}$$

$w_i = (t_i - t_{i-1})\lambda \qquad t_0 \equiv z_0 \equiv 0$

Condition on calibration masses – divides mass range into intervals.

For $t_i > c_s$ use (4).

For $c_j < t_{h+1}, \ldots, t_{h+n} < c_{j+1}$: Dirichlet

$$\frac{\prod_{i=1}^{n+1} \tilde{z}_i^{w_i-1} \Gamma(\sum_{i=1}^{n+1} w_i)}{\prod_{i=1}^{n+1} \Gamma(w_i)}$$

$$\begin{array}{lll} w_i = (t_{h+i} - t_{h+i-1})\lambda & z_h \equiv c_j & T = c_{j+1} - c_j \\ \tilde{z}_i = (z_{h+i} - z_{h+i-1})/T & z_{h+n+1} \equiv c_{j+1} & \end{array}$$

(Does not allow obs. error).

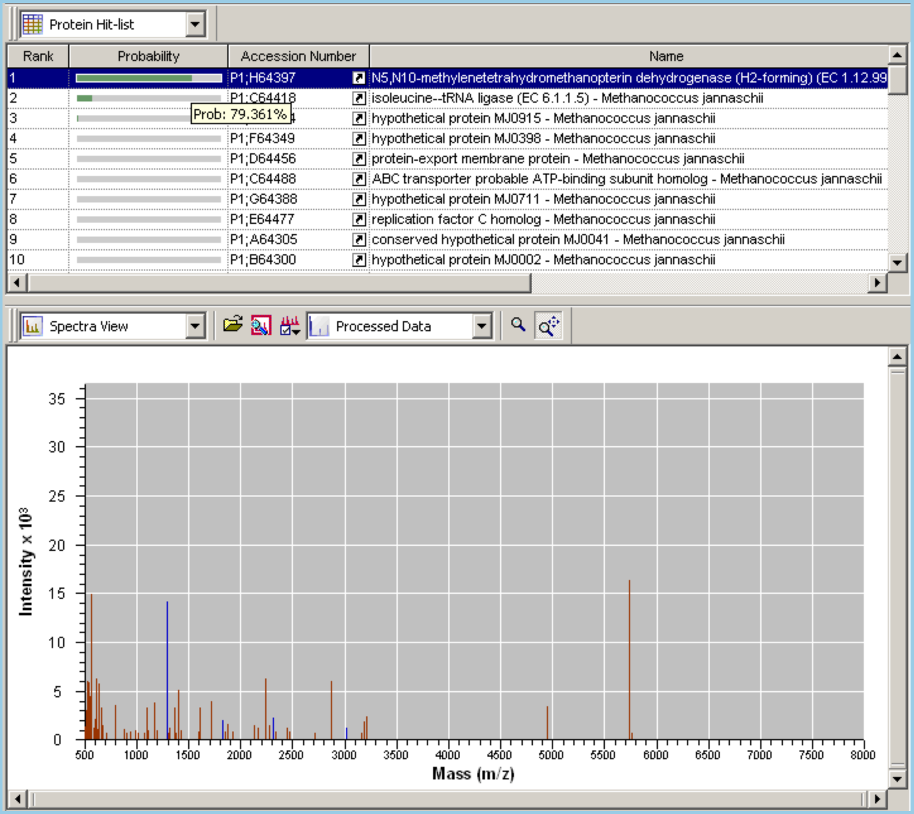
“Extras”

$$\Pr(P_y \mid k_i = k^*, a_{yij}, \lambda_q) = e^{-\lambda_q r} \lambda_q^{q_{yij}}. \tag{5}$$

$$\lambda_q \sim \text{gamma}(a_q, b_q)$$

$$\Pr(P_y \mid k_i = k^*, a_{yij}) = \frac{\Gamma(a_q + q_{yij})}{\Gamma(a_q)} \frac{b_q^{q_{yij}}}{(b_q + r)^{a_q + q_{yij}}}.$$

Results display.

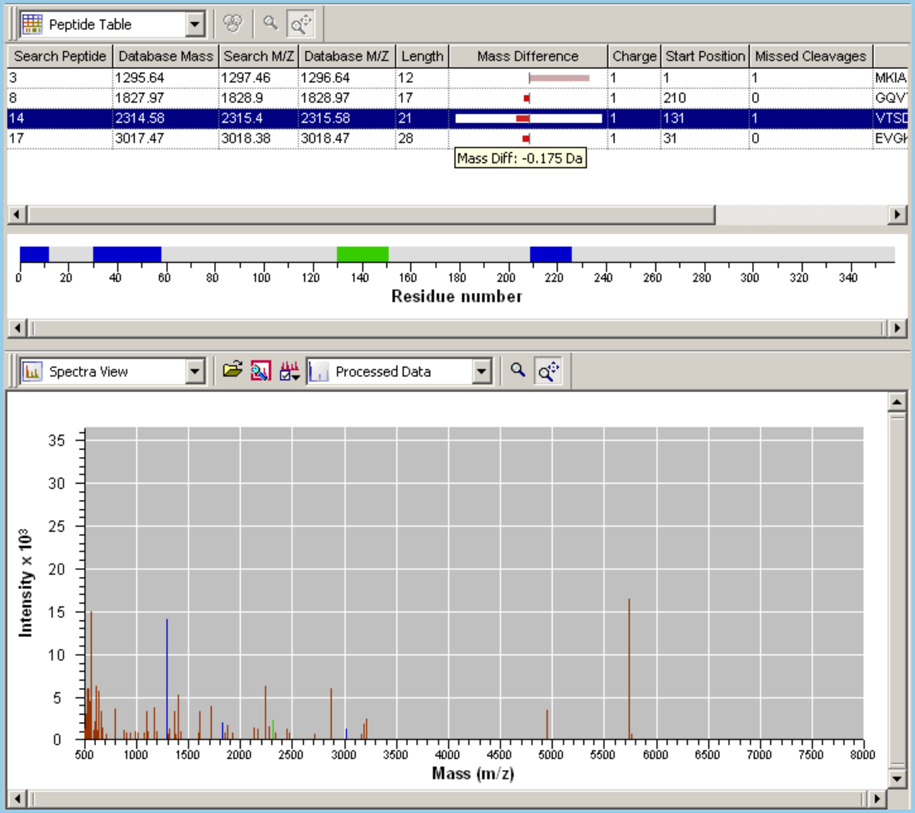


Protein Hit-list

Rank	Probability	Accession Number	Name
1		P1;H64397	N5,N10-methylenetetrahydromethanopterin dehydrogenase (H2-forming) (EC 1.12.99
2		P1;C64418	isoleucine--tRNA ligase (EC 6.1.1.5) - Methanococcus jannaschii
3		[illegible]4	hypothetical protein MJ0915 - Methanococcus jannaschii
4		P1;F64349	hypothetical protein MJ0398 - Methanococcus jannaschii
5		P1;D64456	protein-export membrane protein - Methanococcus jannaschii
6		P1;C64488	ABC transporter probable ATP-binding subunit homolog - Methanococcus jannaschii
7		P1;G64388	hypothetical protein MJ0711 - Methanococcus jannaschii
8		P1;E64477	replication factor C homolog - Methanococcus jannaschii
9		P1;A64305	conserved hypothetical protein MJ0041 - Methanococcus jannaschii
10		P1;B64300	hypothetical protein MJ0002 - Methanococcus jannaschii

Prob: 79.361%

Results display.



Peptide Table

Search Peptide	Database Mass	Search M/Z	Database M/Z	Length	Mass Difference	Charge	Start Position	Missed Cleavages	
3	1295.64	1297.46	1296.64	12		1	1	1	MKIA
8	1827.97	1828.9	1828.97	17		1	210	0	GQV
14	2314.58	2315.4	2315.58	21		1	131	1	VTSD
17	3017.47	3018.38	3018.47	28		1	31	0	EVGK

Mass Diff: -0.175 Da

Spectra View

Processed Data

“Does it work?”

- Bayesian inference.
- Is our belief structure *valid*?
  - “model”
  - “prior”
  - calibration

  Diagnostic checking – prior predictive distribution.
- Are we using all of the available information?

The "glass slipper"

What if $k^* \notin S$?

Simple normal example (e.g. slipper sizes).

Collection of "items" with "true values" $X_i \sim N(\mu, \sigma_X^2)$.

Sample from item i. Observe Y_i where

$$Y_i | X_i \sim N(X_i, \sigma_Y^2).$$

Find a new, unknown, item, with unknown X, then

$$Y \sim N(\mu, \sigma_X^2 + \sigma_Y^2).$$

Assume, for now, that we know the values of μ, σ_X^2, σ_Y^2.

Database containing n known X values, $x_1, \ldots, x_n$ and we suppose that there are m other items not in the database. We observe a sample from an unidentified item and make the observation y.

Prior probabilities $p_1^{(0)}, \ldots, p_n^{(0)}$ for items in the database and $p_{n+1}^{(0)}, \ldots, p_{n+m}^{(0)}$ for items not in the database.

$$f(y, a, b) = \frac{1}{\sqrt{2\pi b}} \exp\left\{ -\frac{1}{2b}(y - a)^2 \right\}.$$

Then posterior probabilities, for $1 \leq i \leq n$,

$$p_i^{(1)} \propto k_i = p_i^{(0)} f(y, x_i, \sigma_{\mathrm{Y}}^2).$$

and, for $n + 1 \leq i \leq n + m$,

$$p_i^{(1)} \propto k_i = p_i^{(0)} f(y, \mu, \sigma_X^2 + \sigma_Y^2).$$

Posterior probability that the sample came from an item not in the database, i.e. *any* item not in the database, is

$$P_O^1 = \frac{\sum_{j=n+1}^{m} k_j}{\sum_{j=1}^{n} k_j + \sum_{j=n+1}^{m} k_j}.$$

Suppose $p_1^{(0)} = \cdots = p_n^{(0)} = p_0^{(0)}$ and $p_{n+1}^{(0)} = \cdots = p_{n+m}^{(0)} = Qp_0^{(0)}$.

Now $np_0^{(0)} + mQp_0^{(0)} = 1$ so

$$Q = \frac{1 - np_0^{(0)}}{mp_0^{(0)}} = \frac{P_O^{(0)}}{mp_0^{(0)}}$$

where

$$P_O^{(0)} = \frac{Qm}{n + Qm} = 1 - np_0^{(0)} = 1 - P_I^{(0)}$$

The posterior probability that the sample came from an item not in the database becomes

$$P_O^{(1)} = \frac{P_O^{(0)} f(y, \mu, \sigma_X^2 + \sigma_Y^2)}{P_I^{(0)} \bar{f}(y, x_1, \ldots, x_n, \sigma_Y^2) + P_O^{(0)} f(y, \mu, \sigma_X^2 + \sigma_Y^2)}$$

where

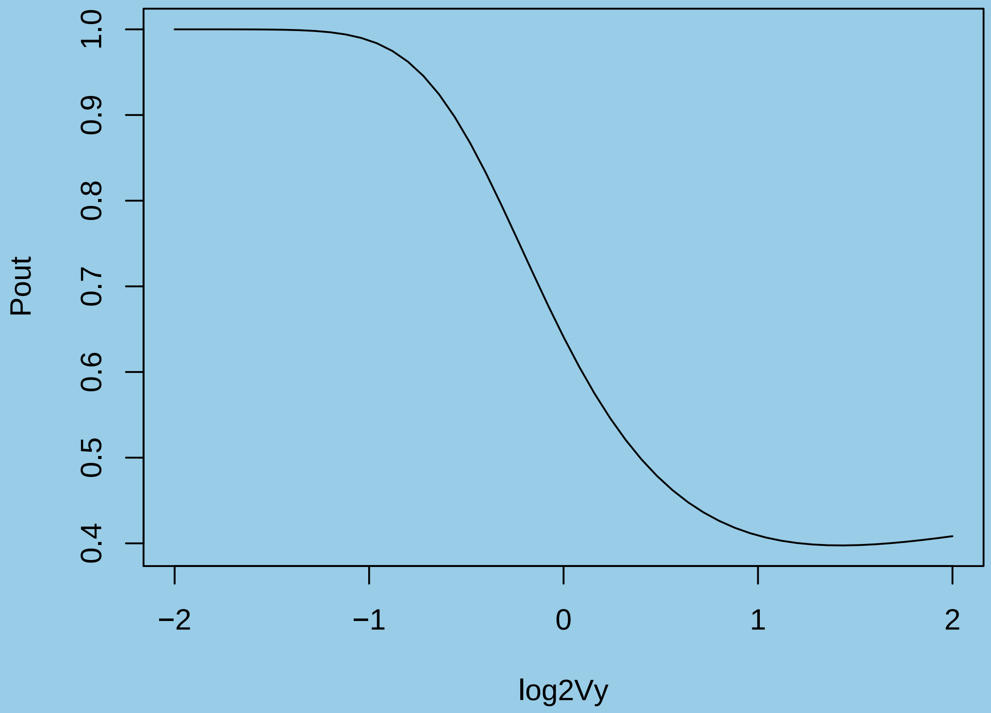
$$\bar{f}(y, x_1, \ldots, x_n, \sigma_Y^2) = \frac{1}{n} \sum_{i=1}^{n} f(y, x_i, \sigma_Y^2).$$

Example: Database, size 10

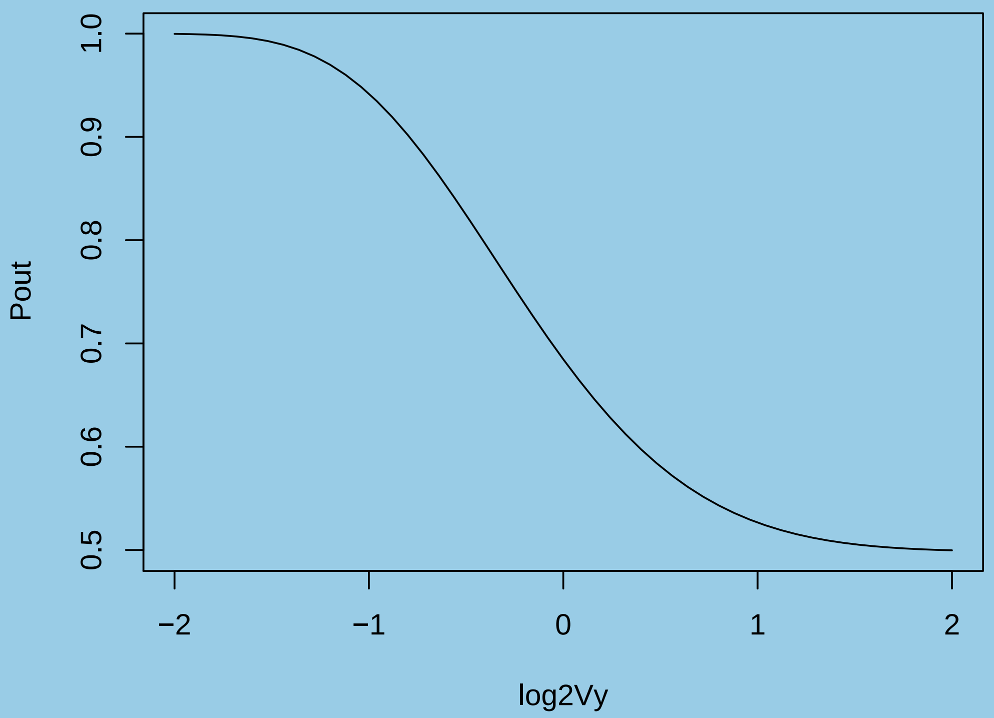
“Database” values:

-0.61 1.64 -0.38 -0.70 -0.13
0.28 1.43 1.27 1.38 0.55

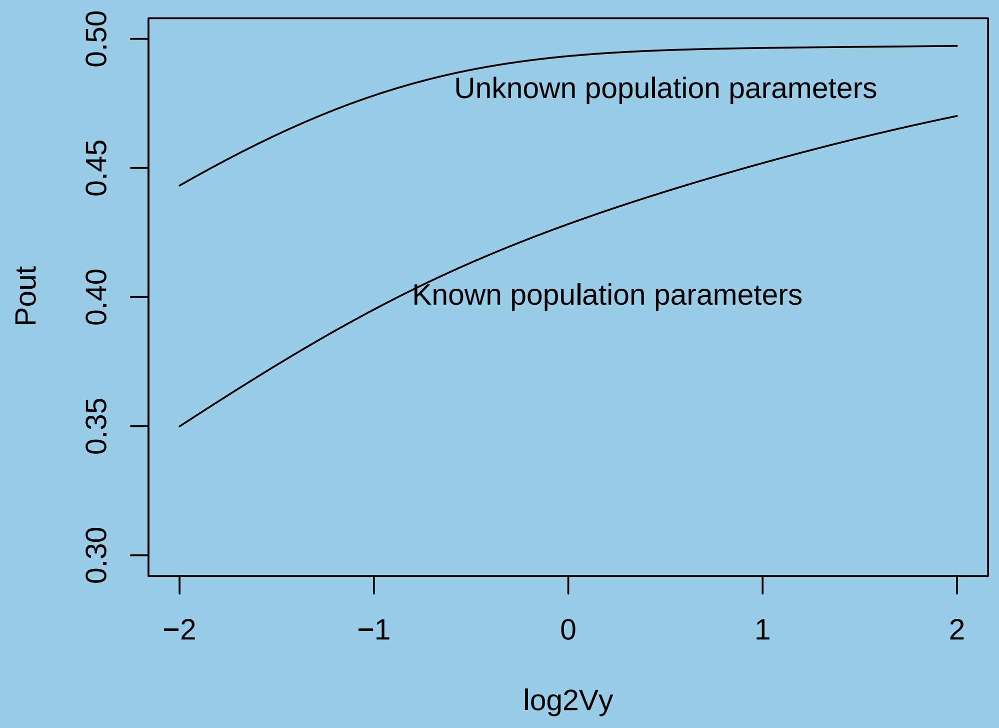
$$P_O^{(0)} = P_I^{(0)} = 0.5$$



Example: $P_O^{(1)}$ against $\log_2(\sigma_Y^2)$, $\sigma_X^2 = 1$, $\mu = 0$, $y = 5$.



$P_O^{(1)}$ against $\log_2(\sigma_Y^2)$, $\sigma_X^2 \sim \mathrm{IG}(1, 0.1)$, $\mu \sim N(0, 10)$, $y = 5$.



Example: $P_O^{(1)}$ against $\log_2(\sigma_Y^2)$, $y = 1.3$.

Application to protein matching

- Mass range, width r.
- Within range, masses from “unknown” protein: Poisson process with rate λ_u.
- Each of these masses has, independently, a probability π of appearing as a peak.

Probability (density) that such an “unknown” protein would give rise to n_a observed peaks in particular locations is

$$e^{-\lambda_u \pi r}(\lambda_u \pi)^{n_a}.$$

Additional peaks in P_y must be “extras”. Use (5) and sum over possible allocations. Obtain a “likelihood” for an “unknown protein”

$$f_u(y, \lambda_u, \pi, \lambda_q, r) =$$

$$\sum_{n_a=0}^{n_y} \left(\begin{array}{c} n_y \\ n_a \end{array} \right) e^{-\lambda_u \pi r} (\lambda_u \pi)^{n_a} e^{\lambda_q \tau} \lambda_q^{n_y - n_a}$$

Allow for uncertainty in $\lambda_u \pi$ and in λ_q :

$$\begin{aligned} \lambda_u \pi &\sim \mathrm{gamma}(a_u, b_u) \\ \lambda_q &\sim \mathrm{gamma}(a_q, b_q) \end{aligned}$$

$$L_{yo} = \sum_{n_a=0}^{n_y} \left(\begin{array}{c} n_y \\ n_a \end{array} \right) G_u(n_a) G_q(n_a)$$

where

$$\begin{aligned} G_u(n_a) &= \frac{\Gamma(a_u + n_a)}{\Gamma(a_u)} \frac{b_u^{a_u}}{(b_u + r)^{a_u + n_a}} \\ G_q(n_a) &= \frac{\Gamma(a_q + n_y - n_a)}{\Gamma(a_q)} \frac{b_q^{a_q}}{(b_q + r)^{a_q + n_y - n_a}} \end{aligned}$$

Imagine prior probabilities $p^{(0)}_{n+1}, \ldots, p^{(0)}_{n+m}$ for the m hypothetical "unknown" proteins and suppose that these are not associated with different beliefs about $\lambda_u \pi$.

Let

$$\mathrm{E}\left(\sum_{j=1}^{m} p^{(0)}_{n+j}\right) = P^{(0)}_0.$$

Then

$$\begin{aligned}\Pr(k^* = k_i|y) &= p_i^{(1)} \\ &= \frac{p_i^{(0)} L_{yi}}{\sum_{j=1}^{n} p_j^{(0)} L_{yj} + P_0^{(0)} L_{yo}}.\end{aligned}$$

Acknowledgments

This work was supported by the **Teaching Company Scheme** in a programme involving the *University of Sunderland* and *Nonlinear Dynamics Ltd.*

The work was connected with the development of a MS data analysis module for Nonlinear Dynamics' software

Progenesis Informatics.