Формулы для оценки качества групп специфичности (т.е. задача такая: есть выравнивание разбитое на группы специфичности (и соответственно есть SDP) надо сказать насколько хороши группы, надо чтоб мера сравнимой для разных выравниваний). Для каждой группы специфичности считается средняя взаимная энтропия по SDP:

$$S_{gr}=\frac{1}{n_{SDP}}\times\sum_{p\in SDP}\sum_{\alpha=1}^{20} f_{p,gr}(\alpha)\ln(\frac{f_{p,gr}(\alpha)}{g_p(\alpha)})$$

где:

n_{SDP} — количество SDP

α — аминокислота

$f_{p,gr}(\alpha)$ — частота остатка α в позиции p в группе gr

$g_p(\alpha)$ — частота остатка α в позиции p во всем выравнивании

Z-score:

$$Z-score(S_{gr})=\frac{S-M(S)}{\sqrt{D(S)}}$$

где:

M(S) — матожидание

D(S) — дисперсия

Z-score считается в следующей модели:

1. SDP — это выборка с возвращением n_{SDP} позиций из всех позиций выравнивания
2. Группа специфичности — это выборка без возвращения из всех последовательностей выравнивания. Объем группы постоянен (равен объему данной группы специфичности)

M(S) и D(S) считал теоретически:

Введем обозначения:

l — длина выравнивания

n — толщина выравнивания

α, β — аминокислоты

$n_p(\alpha)$ — количество аминокислоты α в позиции p во всем выравнивании

$n_{gr}(\alpha)$ — количество аминокислоты α в позиции p в данной группе специфичности (в данной выборке последовательностей)

n_{SDP} — количество SDP

$f_{p,gr}(\alpha)$ — частота остатка α в позиции p в группе gr

$g_p(\alpha)$ — частота остатка α в позиции p во всем выравнивании

1. Рассмотрим случайную величину $S_{p,s}$ — взаимная энтропия группы объема s в столбце p. Вероятностное пространство представляет из себя все возможности выбора без возвращения s последовательностей из n.
2. Вычислим $M(S_{p,s})$:

$$M(S_{p,s})=M\left(\sum_{\alpha=1}^{20} f_{p,gr}(\alpha)\ln(\frac{f_{p,gr}(\alpha)}{g_p(\alpha)})\right)=\sum_{\alpha=1}^{20} M\left(f_{p,gr}(\alpha)\ln(\frac{f_{p,gr}(\alpha)}{g_p(\alpha)})\right)=$$

$$\sum_{\alpha=1}^{20}\sum_{n_{gr}(\alpha)=max(0,\,s+n_p(\alpha)-n)}^{min(s,n_p(\alpha))}\left(p_{n,s,n_p(\alpha)}(n_{gr}(\alpha))\times\frac{n_{gr}(\alpha)}{s}\ln(\frac{n_{gr}(\alpha)/s}{n_p(\alpha)/n})\right)$$

где $p_{n,s,n_p(\alpha)}(n_{gr}(\alpha))=\frac{C_{n_p(\alpha)}^{n_{gr}(\alpha)}\times C_{n-n_p(\alpha)}^{s-n_{gr}(\alpha)}}{C_n^s}$ - вероятность получить $n_{gr}(\alpha)$ остатков α при выборе без возвращения s последовательностей из n при условии, что из них $n_p(\alpha)$ содержат остаток α.

3. Вычислим $D(S_{p,s})$:

$$D(S_{p,s})=M\left(\sum_{\alpha=1}^{20} f_{p,gr}(\alpha)\ln(\frac{f_{p,gr}(\alpha)}{g_p(\alpha)})\right)^2-M(S_{p,s})^2;$$

$$M\left(\sum_{\alpha=1}^{20} f_{p,gr}(\alpha)\ln(\frac{f_{p,gr}(\alpha)}{g_p(\alpha)})\right)^2=\sum_{\alpha=1}^{20}\sum_{\beta=1}^{20} M\left(f_{p,gr}(\alpha)f_{p,gr}(\beta)\ln(\frac{f_{p,gr}(\alpha)}{g_p(\alpha)})\ln(\frac{f_{p,gr}(\beta)}{g_p(\beta)})\right)=$$

$$\sum_{\alpha=1}^{20}\sum_{n_{gr}(\alpha)=max(0,s+n_p(\alpha)-n)}^{min(s,n_p(\alpha))}\left(p_{n,s,n_p(\alpha)}(n_{gr}(\alpha))\times\frac{n_{gr}(\alpha)^2}{s}\ln^2(\frac{n_{gr}(\alpha)/s}{n_p(\alpha)/n})\right)+$$

$$\sum_{\alpha=1}^{20}\sum_{\beta=1}^{20}\left(\sum_{n_{gr}(\alpha)=max(0,s+n_p(\alpha)-n)}^{min(s,n_p(\alpha))}\ \sum_{n_{gr}(\beta)=max(0,s+n_p(\beta)+n_p(\alpha)-n-n_{gr}(\alpha))}^{min(s-n_{gr}(\alpha),n_p(\beta))}\left\{D_{n,s,n_p(\alpha),n_{gr}(\beta)}(n_{gr}(\alpha),n_p(\beta))\right\}\right);$$

где:

$$D_{n,s,n_p(\alpha),n_{gr}(\beta)}(n_{gr}(\alpha),n_p(\beta))=p_{n,s,n_p(\alpha),n_{gr}(\beta)}(n_{gr}(\alpha),n_p(\beta))\times$$
$$\left(\frac{n_{gr}(\alpha)}{s}\ln(\frac{n_{gr}(\alpha)/s}{n_p(\alpha)/n})\right)\times\left(\frac{n_{gr}(\beta)}{s}\ln(\frac{n_{gr}(\beta)/s}{n_p(\beta)/n})\right);$$

где: $p_{n,s,n_p(\alpha),n_{gr}(\beta)}(n_{gr}(\alpha),n_p(\beta))=\frac{C_{n_p(\alpha)}^{n_{gr}(\alpha)}\times C_{n_p(\beta)}^{n_{gr}(\beta)}\times C_{n-n_p(\alpha)-n_p(\beta)}^{s-n_{gr}(\alpha)-n_{gr}(\beta)}}{C_n^s}$ -вероятность получить $n_{gr}(\alpha)$ остатков α и $n_{gr}(\beta)$ остатков β при выборе без возвращения s последовательностей из n при условии, что из них $n_p(\alpha)$ содержат остаток α и $n_p(\beta)$ содержат остаток β.

4. Введем случайную величину $\vec{n}=\{n_i\}_{i=1}^{i=l}$ которая принимает значения $\vec{n}_j=\{0,\ldots,0,n_j=1,0,\ldots,0\}$ с вероятностью 1/l.
5. Рассмотрим случайную величину S_s равную скалярному произведению:
   $S_s=\langle\vec{n}|\{S_{p,s}\}_{p=1}^{p=l}\rangle$ - т.е. случайная величина равновероятно являющейся энтропией любого столбца.
6. Вычислим матожидание S_s:

$$M(S_s)=\sum_{p=1}^{p=l}\frac{1}{l}\times M(S_{p,s})=M(M(S_{p,s}))$$

7. Вычислим дисперсию S_s:

$$D(S_s)=M(S_s^2)-M^2S_s=\frac{1}{l}\sum_{p=1}^{l}M(S_{p,s}^2)-\frac{1}{l}\sum_{p=1}^{l}M^2(S_{p,s})+\frac{1}{l}\sum_{p=1}^{l}M^2(S_{p,s})-M^2(S_s)=$$
$$M(D(S_{p,s}))+D(M(S_{p,s}))$$

8. Т.к. S_s зависимы (поскольку выбор последовательностей в группу специфичности один для всех позиций) то вычислим ковариацию $cov(S_s^1,S_s^2)$ — где S_s^1 и S^2_s величины распределенные как S_s:

$$cov(S_s^1,S_s^2)=M((S_s^1-M(S_s^1))\times(S_s^2-M(S_s^2)))=M(S_s^1S_s^2)-M^2(S_s);$$

$$M(S_s^1S_s^2)=\frac{1}{l^2}\sum_{p_1=1}^{l}\sum_{p_2=1}^{l}M(S_{p_1,s}S_{p_2,s})=$$

$$\frac{1}{l^2}\sum_{p_1=1}^{l}\sum_{p_2=1}^{l}\sum_{\alpha=1}^{20}\sum_{\beta=1}^{20}M\left\{\left(f_{p,gr}(\alpha)\ln(\frac{f_{p,gr}(\alpha)}{g_p(\alpha)})\right)\times\left(f_{p,gr}(\beta)\ln(\frac{f_{p,gr}(\beta)}{g_p(\beta)})\right)\right\}$$

подробностей вычисления последнего матожидания не привожу. Суть простая — все последовательности развиваются на 4 класса:

1. В позиции p_1 находится α в позиции p_2 находится β

2. В позиции p_1 находится α в позиции p_2 находится не β
3. В позиции p_1 находится не α в позиции p_2 находится β
4. В позиции p_1 находится не α в позиции p_2 находится не β
   Далее перебираются все допустимые значения $n_{gr}(\alpha)$ и $n_{gr}(\beta)$ и все способы их реализации при помощи различных комбинаций последовательностей 4-х классов. Вероятность считается аналогично предыдущим (произведение числа способов выбрать нужное число последовательностей для всех классов делить на C^k_n)

9. Наша энтропия является средним по n_{SDP} S^p_s:

$$S_{gr} = \frac{1}{n_{SDP}} \times \sum_{p=1}^{n_{SDP}} S_s^p$$

ее матожидание очевидно равно $M(S_s)$ а дисперсия равна:

$$D(S_{gr}) = \frac{1}{n_{sdp}^2} \sum_{p=1}^{n_{SDP}} D(S_s^p) + \frac{2}{n_{sdp}^2} \sum_{p1=1}^{n_{SDP}} \sum_{p2=p1+1}^{n_{sdp}} cov(S_s^{p1}, S_s^{p2}) = \frac{1}{n_{SDP}} D(S_s) + \frac{n_{SDP}-1}{n_{SDP}} cov(S_s^1, S_s^2)$$