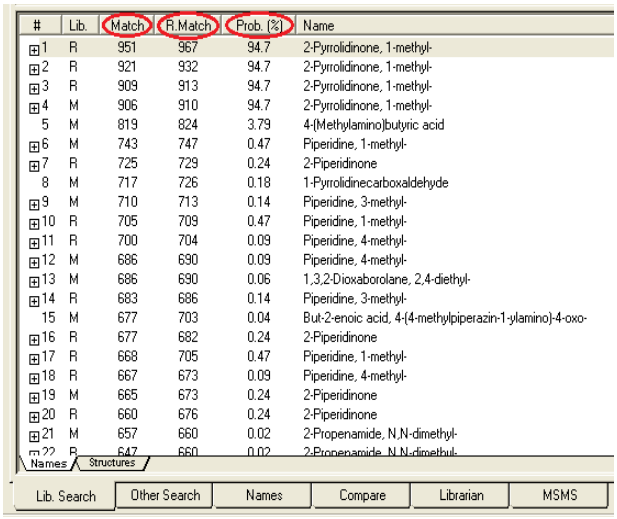
NIST/EPA/NIH Mass Spectral Library Compound Scoring: Match Factor, Reverse Match Factor, and Probability

What is NIST/EPA/NIH Mass Spectral Library?

The National Institute of Standards and Technology, or NIST, is one of the nation's oldest physical science laboratories founded in 1901. This organization has collected electron ionization (EI) gas chromatography (GC) mass spectral data of known standards from various sources to create a mass spectral reference library of compounds. This library is used all over the world for the identifications of unknowns in GCMS chromatograms.

How does it identify unknowns?

NIST uses the submitted unknown spectrum collected by mass spectrometer detectors (MSD) and performs a library spectrum search. The mass spectral data is analyzed against all the library spectra and calculates three numbers associated with each compound. The three numbers are: Match Factor (Match), Reverse Match Factor (R. Match), and Probability (%). These numbers are used to create a hit list of compounds that NIST has identified as possible matching chemical structures, as shown in **Figure 1**.



#	Lib.	Match	R.Match	Prob. (%)	Name
1	R	951	967	94.7	2-Pyrrolidinone, 1-methyl-
2	R	921	932	94.7	2-Pyrrolidinone, 1-methyl-
3	R	909	913	94.7	2-Pyrrolidinone, 1-methyl-
4	M	906	910	94.7	2-Pyrrolidinone, 1-methyl-
5	M	819	824	3.79	4-(Methylamino)butyric acid
6	M	743	747	0.47	Piperidine, 1-methyl-
7	R	725	729	0.24	2-Piperidinone
8	M	717	726	0.18	1-Pyrrolidinecarboxaldehyde
9	M	710	713	0.14	Piperidine, 3-methyl-
10	R	705	709	0.47	Piperidine, 1-methyl-
11	R	700	704	0.09	Piperidine, 4-methyl-
12	M	686	690	0.09	Piperidine, 4-methyl-
13	M	686	690	0.06	1,3,2-Dioxaborolane, 2,4-diethyl-
14	R	683	686	0.14	Piperidine, 3-methyl-
15	M	677	703	0.04	But-2-enoic acid, 4-[4-methylpiperazin-1-ylamino]-4-oxo-
16	R	677	682	0.24	2-Piperidinone
17	R	668	705	0.47	Piperidine, 1-methyl-
18	R	667	673	0.09	Piperidine, 4-methyl-
19	M	665	673	0.24	2-Piperidinone
20	R	660	676	0.24	2-Piperidinone
21	M	657	660	0.02	2-Propenamide, N,N-dimethyl-
22	R	647	660	0.02	2-Propenamide, N,N-dimethyl-

Figure 1: *Example NIST Hit List.*

Match Factor

A match factor, or a direct match, is a comparison of the unknown's mass spectrum's peaks to those of the peaks in the library's spectra. This number therefore is an indication of how similar the unknown's spectrum is to the library's known's spectrum. NIST uses a scale of 0-1000, but no compound score can exceed 999. A score of 999 is a perfect match with all peaks in both spectra matching, while a score of 0 indicates that no peaks in the unknown are in common with the known compound's spectrum. NIST's suggested general guidelines for Match Factor scores are: >900 is an Excellent Match, 800-900 is a Good Match, 700-800 is a Fair Match, and <600 is a poor match.

Reverse Match Factor

A reverse match factor is the match factor when the peaks in the unknown's spectrum that are not in the library's known reference spectrum are ignored. This number is useful when two compounds are co-eluting in the chromatogram. In **Figure 2** middle panels, displayed is the difference between an unknown and reference spectra. The left panel shows the difference between the unknown and the top hit of the hit list, while the right panel shows the difference between the unknown and the fifth hit of the hit list. Red Ions are those that are

different in the unknown spectrum and are ignored when calculating the reverse match score. The blue and green ions are the major influence in the reverse match score. In the left panel all ion differences are minor, while those in the right panel are more significant. This is the reason why the R. Match for the left is 967, while the right is 824.

Probability

Probability for NIST is determined by first assuming that the unknown's spectrum is present in the library. It then uses the hit list generated from the match factors to compare the adjacent hits in the hit list taking into account the differences between these hits to determine a relative probability that any of the hits in the hit list are correct. For a hit list, if the first hit has a high match factor (>900) while the next hit in the hit list instead has a significantly lower match factor (≤800), then for that compound the probability of it being identified correctly is high and the likelihood of it being the library is also high. See **Figure 2** for an example of >900 match factors vs. ~800 match factors, and how probability is affected.

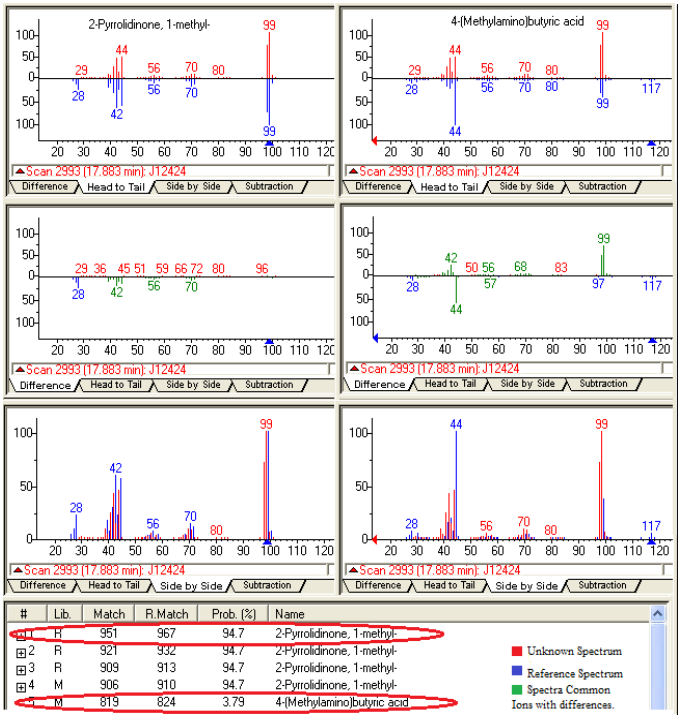


Figure 2: *Top Hit vs. Fifth Hit in Hit List. Top Panels: Head-to-Tail. Middle Panels: Difference. Bottom Panels: Side-by-Side*

What can lower these numbers for a correct compound?

- Match Factors scores are affected by how many peaks are in a spectrum. Unknown spectra that have many peaks in it tend to give a lower score than similar spectra that have fewer peaks. The larger the number of peaks present in an unknown's spectrum, the greater number of peaks that have to match to the reference spectrum to receive a high match factor. This can be caused by a high and/or noisy baseline, as well as co-elution. The reverse match factor can still be high in these cases as it ignores all non-matching peaks for its comparison of unknown and reference spectra.
- Probability is affected when many compounds have very similar mass spectra to the unknown spectrum. When the unknown spectrum has a limited number of spectra it is similar to, it will have a higher probability value then if it had a spectrum that was consistent with many other mass spectra.

Due to the above problems, well matched spectra may show poor scores and poorly matched spectra may show good scores. It is therefore important for an experienced chemist to verify results from the database for questionable compounds. Knowledge of the sample chemistry can be used to refine the database identifications.

https://www.nist.gov/sites/default/files/documents/srd/NIST1aVer22Man.pdf