Measuring peptide mass spectrum correlation using the quantum Grover algorithm

Keng Wah Choo*

Bioinformatics Group, Nanyang Polytechnic, 180 Ang Mo Kio Avenue 8, Singapore 569830, Singapore
and National Institute of Education, Nanyang Technological University, 1, Nanyang Walk, Singapore 637616, Singapore

(Received 16 August 2006; revised manuscript received 22 December 2006; published 30 March 2007)

We investigated the use of the quantum Grover algorithm in the mass-spectrometry-based protein identification process. The approach coded the mass spectra on a quantum register and uses the Grover search algorithm for searching multiple solutions to find matches from a database. Measurement of the fidelity between the input and final states was used to quantify the similarity between the experimental and theoretical spectra. The optimal number of iteration is proven to be $\frac{\pi}{4}\sqrt{\frac{N}{k}}$, where k refers to the number of marked states. We found that one iteration is sufficient for the search if we let more that 62% of the N states be marked states. By measuring the fidelity after only one iteration of Grover search, we discovered that it resembles that of the correlation-based measurement used in the existing protein identification software. We concluded that the quantum Grover algorithm can be adapted for a correlation-based mass spectra database search, provided that decoherence can be kept to a minimum.

DOI: 10.1103/PhysRevE.75.031919 PACS number(s): 87.14.Ee, 87.10.+e, 03.67.Lx

I. INTRODUCTION

The invention of the Grover search algorithm [1,2] gives rise to many potential speed-up applications offered by a quantum computer [3]. Its theoretical analyses and practical implementations have been the focus of research since then. A variety of applications were developed based on some variations of the search algorithm [4–6] to solve classical problems. We are interested in applying the Grover algorithm to the field of proteomics, in particular, the search of matching mass spectra from a large collection of theoretical mass spectra of short peptides, calculated from known protein sequences.

A. Grover search algorithm

We consider a search space D containing N elements, assuming that $N=2^n$, where n is an integer. The elements of D are represented using an n-qubit register containing the indices $i=0,1,\ldots,N-1$. We assume that there are k marked elements in the search space that are the solution to the search problem. It is also assumed that there exists a function $f:D \rightarrow 0,1$, such that $f=1$ for the marked elements and $f=0$ for the rest. The search for a marked element becomes a search for an element for which $f=1$. To solve this problem using a classical computer, one needs to evaluate f for each element, one by one, until a marked state is found. Thus on average, $N/2$ evaluations of f are necessary and in the worst case, N evaluations are required. For a quantum computer, the function f can be evaluated coherently. The Grover search algorithm uses a sequence of unitary operations and it can locate a marked element using only $O(\sqrt{N})$ coherent queries of f, which is faster than classical computers. In the case of k marked states exist in the database, the number of optimal iterations is given as $r_o=\frac{\pi}{4}\sqrt{\frac{N}{k}}$.

To briefly describe the algorithm, we consider a quantum system having an initial state of register as $|i\rangle=|i_1\cdots i_n\rangle$ of n qubits, the register is subjected to local Hadamard transformations, $H^{\otimes n}$, resulting in a linearly superposed state of all computational basis,

$$|\varrho\rangle = H^{\otimes n}|0\cdots 0\rangle = \frac{1}{\sqrt{N}}\sum_{i=0}^{N-1}|i\rangle, \tag{1}$$

where $H=\frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}$. The states are then subjected to a series of inversion and diffusion processes before measurement. During the inversion stage, the state undergoes the unitary transformation

$$I = 1 - 2\sum_{m_i \in M}|m_i\rangle\langle m_i|, \tag{2}$$

where $M=\{m_1,m_2,\ldots\}$ are the marked states. In the diffusion stage, the state undergoes the transformation

$$D = -1 + 2|\eta\rangle\langle\eta|. \tag{3}$$

In this way, the state prior to measurement after r iterations is $|\psi_r\rangle=(DI)^r|\eta\rangle \equiv U_G{}^r|\eta\rangle$.

B. Mass spectrometry

This section intends to provide some background information to mass spectrometry. Mass spectrometry (MS) is a common analytical technique used to identify unknown compounds, quantify known materials, and elucidate the molecular structure and chemical composition of organic and inorganic substances. A mass spectrometer is an instrument used to measure the mass-to-charge ratio of individual molecules that have been converted into electrically charged molecules, or ions [7]. These ions are filtered and ordered from a lower to higher mass-to-charge ratio (m/z) before passing through an ion detector in the instrument [8]. In the field of proteomic analysis, matrix assisted laser desorption ionization (MALDI) and electrospray ionization (ESI) are two ionization techniques generally used. Mass spectrometry is currently experiencing rapid growth in mass-spectrometry-based biomarker discovery and clinical proteomics, where hundreds of proteins can be sequenced quickly. As a conse-

*Email address: choo_keng_wah@nyp.edu.sg

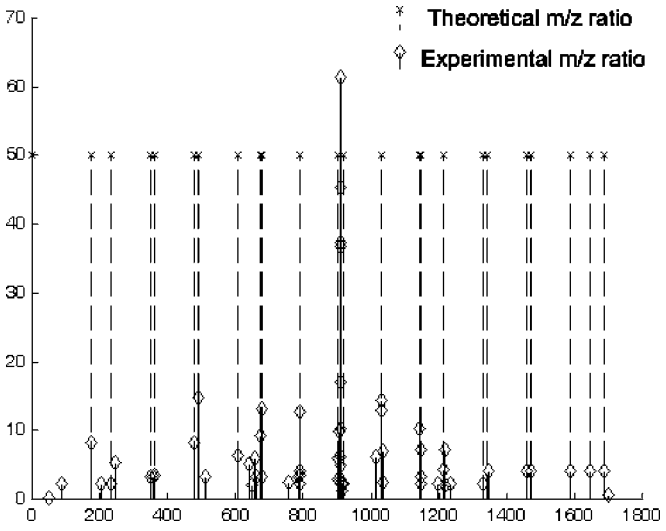

FIG. 1. The overlapping plot of experimental mass spectra against theoretical mass spectra where cross correlation can be computed.

quence, large amounts of proteomics data are produced and made available to the public [9–11]. Although the generation of raw MS spectra has become easier, the analysis and identification of the data is a challenge. Many protein identification tools have been developed, such as PEAKS [12], MASCOT [13], Phenyx [14], and OMSSA [15]. In the case of high throughput proteomics, it involves the analysis of hundreds of thousands of peptide spectra derived from biological samples. These spectra can be identified by four general types of algorithms.

(1) *De novo* calling of the sequence directly from the spectrum [12,16,17].

(2) Use of unambiguous "peptide sequence tags" derived from spectra that are used to search known sequences [18–20].

(3) Cross-correlation methods that correlate experimental spectra with theoretical spectra [21,22].

(4) Probability-based matching that calculates a score based on the statistical significance of a match between an observed peptide fragment and those calculated from a sequence search library [23–27].

Cross-correlation methods and probability-based matching are two well-received methods for protein identification. In these methods, a theoretical mass spectra database is first generated from known protein sequences. To search this database with an experimental spectra, the correlation of the experimental and theoretical spectra is calculated. Based on the statistical properties of the protein database and the correlation values (actual implementation is more complex), a score is given for the matched spectra. Peptide spectra with scores better than a predetermined threshold will be returned as hits. Figure 1 shows the plot of overlapping experimental and theoretical mass spectra.

C. SEQUEST

SEQUEST [28] is one of the search programs that uses a descriptive model for peptide fragmentation and correlative matching to a tandem mass spectrum. It uses a two-tiered scoring scheme to assess the quality of the match between the experimental spectrum and the theoretical ones from a database. The first score calculated, the preliminary score S_p, is an empirically derived score that restricts the number of sequences analyzed in the correlation analysis. The second score is the cross-correlation of the experimental and theoretical spectra. This score is referred to as X_{corr} where the theoretical and normalized experimental spectra are cross-

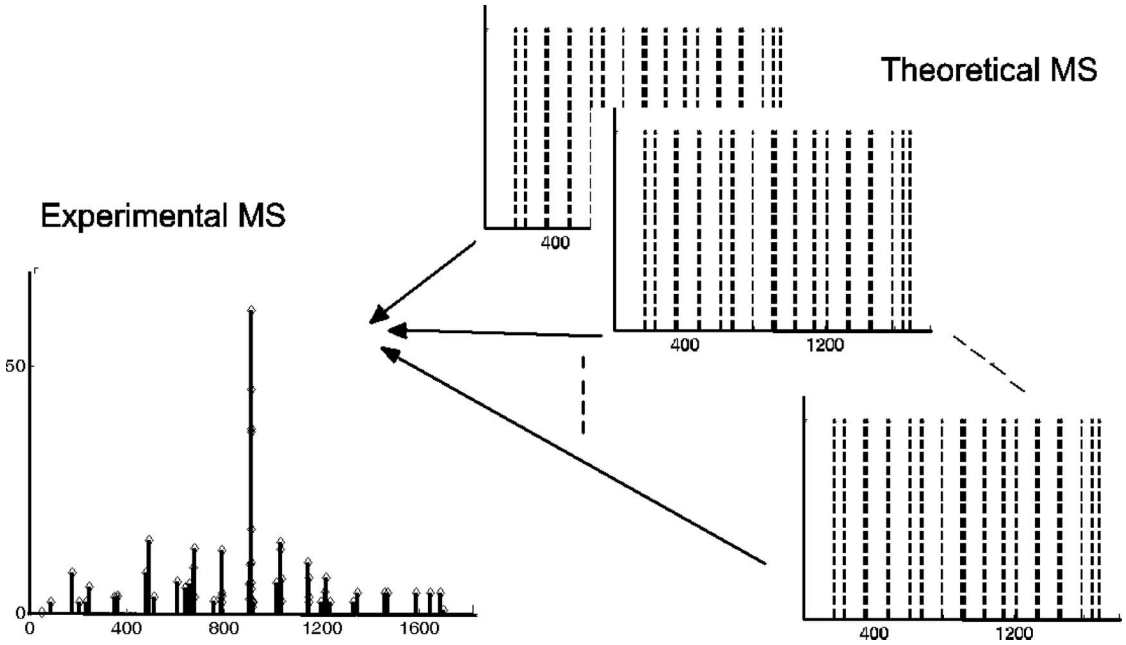

FIG. 2. The conceptual idea of the database search. The experimental mass spectrum is coded as an oracle to be searched by potential theoretical mass spectra.

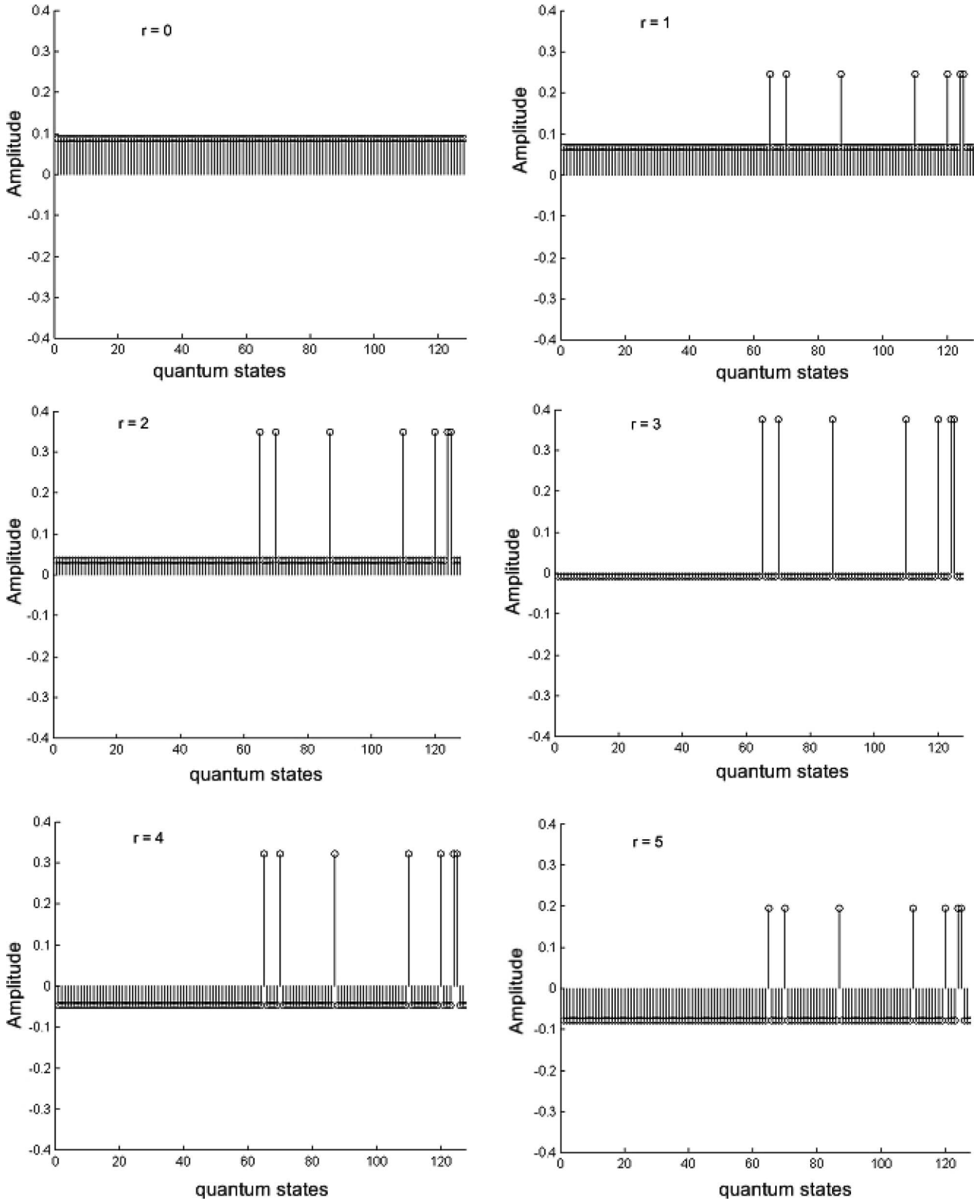

FIG. 3. The plot of quantum states at r iteration with $k=7$ marked states.

correlated to obtain similarities between the spectra, as shown in Eq. (4). The cross-correlation score is further processed with the preliminary score to determine the final score of the spectrum. SEQUEST has been shown to have good sensitivity and flexibility and is applicable to data generated by different types of mass spectrometers:

$$X_{\mathrm{corr}}(E,T) = \sum_{i=0}^{N-1} x_i y_{i+t}. \tag{4}$$

II. METHODS AND MATERIALS

Although correlation-based mass spectra database search algorithms provide fast and accurate protein prediction, implementation of these algorithms in a quantum computer seems challenging. Lomont [29] proved that it is impossible to have a quantum correlation implemented physically on a quantum computer. This is unfortunate for proteomics applications as the opportunity to use a fast quantum algorithm is apparently denied. We propose to use the Grover search algorithm for this purpose. One immediate solution would be to code the experimental mass spectra into a register of quantum states and use it to search an oracle to be generated from all possible theoretical mass spectra. However, this solution has many disadvantages, such as the coding of experimental mass spectra into a quantum register, the coding of theoretical mass spectra into an oracle to be searched, and the complexity of the evaluation function $f(s)$ to be used.

Instead, we propose that the experimental mass spectra be coded as the oracle to be searched, and the theoretical mass spectra be the quantum search states. Figure 2 depicts the conceptual idea of this approach. It can be viewed as a type of Grover search with multiple solutions, i.e., if there exists more than one theoretical m/z value matching that of the experimental one, there will be more than one marked state in the oracle. With the assumption that a typical experimental mass spectrum comprises two types of data, the one generated from actual fragmented ions and that caused by experimental noise, the size of the experimental mass spectrum would always be larger than the theoretical ones. We can then consider the m/z values generated from ions as marked states $|m\rangle$ and the rest (noise) as unmarked states $|m^{\perp}\rangle$. The function $f(s)$ is formed such that if the experimental m/z matches one of the theoretical m/z within an acceptable tolerance, then the function returns 1, otherwise a 0 is returned. The acceptable tolerance depends on the type of mass spectrometry technique used [30] and any increase of mass tolerance above this threshold can potentially increase false positive rates for protein identification as suggested by [25,31]; although one report was found indicating that varying mass tolerance had little effect on the accuracy of protein identification [32].

It is very common for existing algorithms such as Phenyx [14] and OMSSA [15] to perform prefiltering based on the intensity of the peaks. This process aims at removing peaks whose intensities are below a predetermined noise-level threshold. The remaining peaks will be set to equal intensity and used for a database search. Our method assumed that the mass spectrum has gone through the same process, hence all peaks in our simulation are treated as having equal intensity.

Figure 3 shows the quantum states that code the experimental mass spectrum change under several Grover iterations. The process will be stopped when the optimal number of iterations r_o are reached. The probability of success is then measured from the final states, where the sum of the probabilities of the experimental spectra matching those theoretical spectra is calculated. A threshold can then be set by the user whether to accept the search result or reject it. In an ideal case, the probability of success is equal to one at the optimal r_o iterations. Theoretically, the optimal iteration to achieve the best search result is given by $\frac{\pi}{4}\sqrt{\frac{N}{k}}$, where k refers to the number of matched spectra. The fidelity between the input states and the final states can also be calculated using the following formula:

$$F = \langle e|\Psi|e\rangle, \tag{5}$$

where $\Psi = |\psi_r\rangle\langle\psi_r|$ and $|e\rangle$ refers to the search state formed by the theoretical mass spectrum.

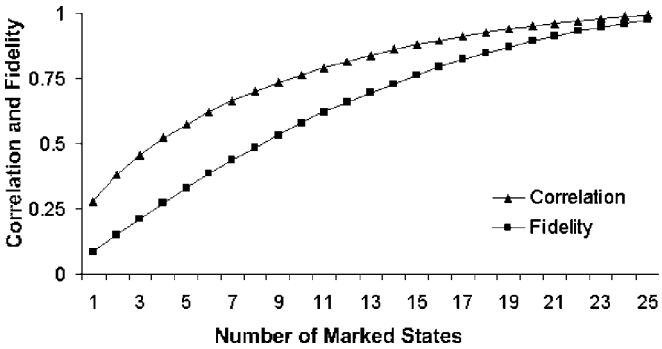

FIG. 4. The plot of correlation and fidelity against the number of marked states k, for $N=64$.

It is interesting to note that the fidelity itself is a possible measure of correlation between the input states and that from the database, as shown in Fig. 4. In other words, the correlation between the experimental mass spectrum and the theoretical mass spectrum can thus be calculated or estimated from the fidelity of the quantum system. In the case when the database is presorted, which can be done on the m/z values, it has been reported in some cases [33–35] where search speed is faster than the classical approach. There is one more problem unsolved: the number of marked states, or the value of k, is unknown *a priori*, hence the optimal number of iteration r_o is unknown. Figure 5 shows the optimal number of iteration r_o computed as the number of solution increases. The number of the solution is the reflection of the number of m/z values matched between the experimental and theoretical spectra. However, since r_o is given by $\frac{\pi}{4}\sqrt{\frac{N}{k}}$, we can safely set a threshold, such that the ratio $\sqrt{\frac{N}{k}}$ is equal to $\frac{4}{\pi}$; then we will only need *one* iteration in all searches having the highest fidelity score. In other words, if 62% $\left(\frac{\pi^2}{16}\times 100\right)$ of the theoretical m/z values match that of the experimental

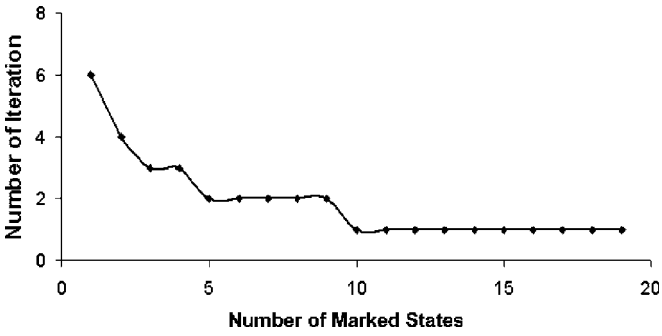

FIG. 5. The plot of optimal number of iteration against the number of marked states k, for $N=64$.

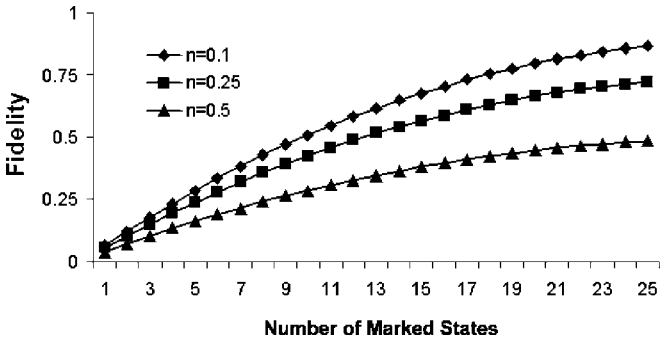

FIG. 6. The plot of fidelity against the number of marked states k, for $N=64$ under the effect of decoherence.

m/z values, the solution can be found in just one iteration, with high fidelity or probability of success.

We investigated the effects of decoherence on this system as all physical implementation of quantum system would have this problem. We considered the noise admixture where the decoherence is achieved by adding a noise admixture coefficient of η into the system after the first Grover iteration. The state of the system is given by Eq. (6). Figure 6 depicts the effect of decoherence on the quantum search result. It is interesting to note that the effect is more significant with a larger number of marked states k, i.e., the system would only be useful under negligible decoherence. Techniques such as error correction or error avoiding can be included to counter the effects of decoherence,

$$\rho_1(t) = (1 - \eta)\rho_1(0) + \frac{\eta}{N} I_N. \tag{6}$$

III. CONCLUSION

This work demonstrated that the famous quantum search algorithm can be adopted for a correlation-based mass spectra search. We considered the experimental spectrum to be coded as the oracle for a search by theoretical mass spectrum, which led to the multiple solution search by the Grover algorithm. Knowing that as the number of solutions increased, the optimal number of iterations reduced. In fact, from the optimal iteration given by $\frac{\pi}{4}\sqrt{\frac{N}{k}}$, if 62% of the experimental m/z values match those from theoretical ones, or $k=0.62N$, only one cycle of search is required. We then showed in our simulation that after one iteration of Grover search, the measurement of the fidelity is sufficient to determine the cross-correlation result of the search. Finally, we found that decoherence can be destructive to the system, hence some forms of error correction of error avoiding mechanisms have to be put in place to ensure the success of the search.

[1] L. K. Grover, Phys. Rev. Lett. **79**, 325 (1997).
[2] L. K. Grover, Phys. Rev. Lett. **79**, 4709 (1997).
[3] S. L. Braunstein and A. K. Pati, Quantum Inf. Comput. **2**, 399 (2002).
[4] B. W. Reichardt and L. K. Grover, Phys. Rev. A **72**, 042326 (2005).
[5] A. Carlini and A. Hosoya, Phys. Lett. A **280**, 114 (2001).
[6] R. Jozsa, e-print quant-ph/9901021.
[7] C. M. Chiu and D. C. Muddiman, What Is Mass Spectrometry?, http://www.asms.org/whatisms
[8] C. G. Herbert and R. A. W. Johnstone, *Mass Spectrometry Basics* (CRC Press LLC, Boca Raton, FL, 2003).
[9] J. V. Puymbrouck, D. Angulo, K. Drew, L. A. Hollenbeck, D. Battre, A. Schilling, D. Jabon, and G. V. Laszewski, DePaul CTI Technical report, 2006 (unpublished).
[10] F. Desiere, E. W. Deutsch, N. L. King, A. I. Nesvizhskii, P. Mallick, J. Eng, S. Chen, J. Eddes, S. N. Loevenich, and R. Aebersold, Nucleic Acids Res. **34**, D655 (2006).
[11] M. Kinter and N. E. Sherman, *Protein Sequencing and Identification Using Mass Spectrometry* (Wiley-Interscience, New York, 2000).
[12] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie, Rapid Commun. Mass Spectrom. **17**, 2327 (2003).
[13] MASCOT by Matrixscience, http://www.matrixscience.com/home.html
[14] Phenyx by Genebio, http://www.phenyx-ms.com/
[15] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant, J. Proteome Res. **3**, 958 (2004).
[16] R. S. Johnson and J. A. Taylor, Mol. Biotechnol. **22**, 301 (2002).
[17] A. Shevchenko, S. Sunyaev, A. Loboda, A. Shevchenko, P. Bork, W. Ens, and K. G. Standing, Anal. Chem. **73**, 1917 (2001).
[18] M. Mann and M. Wilm, Anal. Chem. **66**, 4390 (1994).
[19] S. Sunyaev, A. J. Liska, A. Golod, A. Shevchenko, and A. Shevchenko, Anal. Chem. **75**, 1307 (2003).
[20] D. L. Tabb, S. Saraf, and J. R. Yates III, Anal. Chem. **75**, 6415 (2003).
[21] J. K. Eng, A. L. McCormack, and J. R. Yates III, J. Am. Soc. Mass Spectrom. **5**, 976 (1994).
[22] P. A. Pevzner, V. Dancik, and C. L. Tang, J. Comput. Biol. **7**, 777 (2000).
[23] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell, Electrophoresis **20**, 3551 (1999).
[24] H. I. Field, D. Fenyo, and R. C. Beavis, Proteomics **2**, 36 (2002).
[25] K. R. Clauser, P. Baker, and A. L. Burlingame, Anal. Chem. **71**, 2871 (1999).
[26] D. Fenyo, J. Qin, and B. T. Chait, Electrophoresis **19**, 998 (1998).
[27] N. Zhang, R. Aebersold, and B. Schwikowski, Proteomics **2**, 1406 (2002).
[28] SEQUEST patent: U.S. Patent No. 6,017,693 (25 January 2000).
[29] C. Lomont, e-print quant-ph/0309070.

[30] G. Zhang and T. A. Neubert, Mol. Cell Proteomics **5**, 401 (2006).
[31] J. V. Olsen, S. E. Ong, and M. Mann, Mol. Cell Proteomics **3**, 608 (2004).
[32] D. Fenyo and R. C. Beavis, Anal. Chem. **75**, 768 (2003).
[33] L. K. Grover and J. Radhakrishnan, in *ACM Symposium on Parallel Algorithms and Architectures (SPAA)* (ACM Press, New York, 2005), p. 186; e-print quant-ph/0407122.
[34] H. Buhrman, C. Durr, M. Heiligman, P. Hoyer, F. Magniez, M. Santha, and R. Wolf, SIAM J. Comput. **34**, 1324 (2005).
[35] M. Boyer, G. Brassard, P. Høyer, and A. Tapp, Fortschr. Phys. **46**, 493 (1998).