A fingerprinting Algorithm Based on Artificial Intelligence
Genetic Algorithm

Yong Wang¹, Shaojun Zhang¹*, Ming-yu Wang¹ and Guangyou Xiong¹
¹College of Naval Architecture and Marine Engineering, Shandong Jiaotong University, Weihai; PR China
*Corresponding author e-mail: shaojunyt@163.com

Abstract. Oil spill identification is an important means of investigation of oil spill through analysis and comparison of suspicious oil spill source. This paper proposed a GC/MS fingerprint fitting method based on genetic algorithm (GA) to identify the source of oil spill. In this method, dominant characteristic peaks were selected from the 3d fingerprint of the standard and stored in the prevailing characteristic fingerprint database (dominant database). Recessive characteristic peaks were screened out from the periphery of the dominant characteristic peaks of the sample and stored in the recessive characteristic fingerprint database (recessive database). Then a set of characteristic peaks was selected from the dominant library by GA to correct the characteristic peaks in the samples and automatically identify the similarity between the standard and the samples. Simulation calculation compared with experimental analysis results showed that the method to build a detailed identification of oil spill fingerprint was accurate and reliable. The method to identify the fingerprint peak can be clearly and accurately reflect the properties of the sample, which can be used for rapid automatic calculation of chromatographic fingerprint similarity, so as to strengthen sea spill accident discriminant scientific theory in identification, traceability and the application of legal liability.

Keywords: Oil Spill, Genetic Algorithm, Tracing the Source, 3d Fingerprint Fitting

1. Introduction
Oil spill identification is an important means of evidence collection in oil spill accident investigation and treatment, while oil fingerprint identification is the main technology of oil spill identification at present [1,2]. Petroleum is a complex mixture of hydrocarbons and no n-hydrocarbons with different molecular sizes and chemical structures. Each oil product has molecular characteristics that are obviously different from other oil products. These molecular characteristics with obvious differences are called fingerprints, which are the inherent chemical characteristics of oil products [3,4]. The unique chemical composition of petroleum is the theoretical basis of chemical fingerprint analysis. Oil composition differences related to the production of carbon source and migration environment: extracting the oil is refined through a series of cracking, cracking, fractionation, alkylation, mixing process of formation, refers to the carbon number of light end products in the n-C₃ ~ n-C₁₂ between oil,
gas chromatography (GC) is the main form of low carbon number dissolved hydrocarbon and the
to the UCM (Unresolved Complex Mixture) [5]. The carbon number range of light oil is $n-C_6 \sim n-C_{26}$. For example, GC fingerprinting of diesel oil is mainly composed of the dissolved peak whose
carbon number range is $n-C_{10} \sim n-C_{17}$, with obvious UCM in the middle of the spectrum [6]. Heavy
oils, such as heavy fuel and lubricating oils, each have their own characteristics. The carbon number of
heavy fuel ranges from $n-C_{14}$ to $n-C_{36}$, and contains a large UCM (more than 50% of the GC) [7]. The
oil spectrum is characterized by almost no peaks representing dissolved alkanes and a wide spectrum
(the carbon number range is $n-C_{18}$ to $n-C_{40}$) [8,9]. Because of the composition of oil and oil products
and the proportion are different, which can clearly reflect the oil in the process of generation and
migration of the organic matter input, geological conditions, the maturity of the original information as
well as the harmonic in petroleum products, such as large production process complex information
collection [10,11].

Fingerprint technology mainly includes three parts: extraction, calculation and identification of characteristic fingerprint. Most commonly used methods of fingerprint peak identification are external
and internal marking [12]. External standard method relies on retention time identification and does
not require correction factor, so it is suitable for the determination of samples with simple mass
composition [13]. However, it has high requirements for the stability of operating conditions and the
reproducibility of injection volume, so it is not suitable for the analysis of complex samples. Internal
standard rules need to choose the appropriate internal standard to join, make internal standard
substance and the matching of the substance, and the peak area was greater than 75% [14]. This paper
proposes a GA based on artificial intelligence (AI) chromatographic fingerprint peak matching
recognition method, automatically select the characteristic peak samples used for calibration of
characteristic peak, and the fingerprint peaks of the sample with the standard sample under test pairs
automatically, so as to overcome the inherent defects of artificial selection characteristic peaks,
identify the characteristic peak results are accurate and reliable.

2. Fundamental

GA is a type of stochastic search method based on the evolution law of biology [15,16]. Fingerprinting
can provide information such as retention time and peak area of fingerprint peak. If the fingerprint
peak is described by points, where $T_i$ is the retention time and $A_i$ is the peak area, the fingerprint
can be represented by a set of points. Suppose the point set of the fingerprint to be tested is $P$, and the
point set of the standard fingerprint is $Q$, as showed below:

$$
\begin{align*}
P &= \{(T_1, A_1), (T_2, A_2), \ldots, (T_M, A_M)\} \\
Q &= \{(T_{i1}, A_{i1}), (T_{i2}, A_{i2}), \ldots, (T_{iM}, A_{iM})\}
\end{align*}
$$

Where, $i$ was the characteristic fingerprint peak number of the sample, and $M$ was the
fingerprint peak number of the standard fingerprint. It can be mutually registered with the
corresponding fingerprint peaks in the standard fingerprint. Expansion and translation are shown as
following:

$$
T'_i = a \times T_i + b
$$

Where, $T_i$ was the retention time of the fingerprint peak in the fingerprint to be measured; $T'_i$ the
retention time of the peak after correction; $a$ was the expansion factor; $b$ was the translation factor.

2.1 Establish Recessive Characteristic Peak Bank

There are more than 200 characteristic peaks in the fingerprint of oil samples, and the time complexity
of violent exhaustive search method is too large. When the data are too large, the program runs slowly,
which does not meet the requirements of algorithm complexity and memory space. It is assumed that
the main component peaks in the sample and the standard fingerprint are both 50, and 12 characteristic
peaks should be searched from them, with a complete search space of \( \binom{12}{2} \times \binom{50}{50} \) and a huge amount of computation. In order to reduce the search space of characteristic peaks, reduce the computation and improve the search efficiency, the characteristic adsorption peaks, called \( n \)-strong peaks, can be screened according to the standard fingerprint, and a relatively small recessive database can be established to search out a group of \( n \)-strong peaks.

Characteristic peaks with sharp peak shape, good symmetry and even distribution were selected as \( n \)-strong peaks. Therefore, the standard fingerprint was divided into three parts according to the retention time. There were four strong peaks selected for each part, \( n=12 \). Because of chromatographic fingerprint said better repeatability, the RSD\% < 1\%, distributed in \( n \) strong peak near the standard fingerprint corresponding fingerprint peak position, the reason for this every \( n \) strong peak around the choice of three peak as a recessive characteristic peak, charge for \( T_1 \), \( T_2 \) and \( T_3 \) respectively, into the recessive characteristic peak library STORE, candidate \( n \) strong peaks of the library structure as shown below.

\[
\text{STORE}((x_1, x_2, \ldots, x_{12})) = (T_1, T_2, T_3)
\]

The library contains 12 records, each record contains 5 elements, of which element 1 is the retention time. Elements 2-5 represent the four possible states of the characteristic peak corresponding to the standard fingerprint in the fingerprint to be tested (element 2 represents the absence of the corresponding characteristic peak, let it be 0; element 3, 4 and 5 represent the retention time of the three hidden strong peaks closest to the \( n \) strong peak of the standard fingerprint, respectively).

2.2 GA is Used to Select Characteristic Peaks

2.2.1. Establish Coding Scheme The multi-parameter mapping coding method is used to encode each parameter into sub-strings by binary coding, and then these sub-strings are concatenated into a complete chromosome. There are 10 strong \( n \) peaks, and each characteristic peak has 4 states to be selected, so the chromosome should contain 10 substrings, each of which is encoded with a 2-bit binary string, as showed below:

\[
0 \ 1 \ 1 \ 1 \ 1 \ 0 \ 1 \ 1 \ 1 \ 0 \ \ldots \ \ldots \ x_i \ x_{12} \ x_{13} \ x_{14} \ x_{15} \ x_{16} \ x_{17} \ x_{18} \ x_{19} \ x_{20}
\]

2.2.2. Calculated fitness (1) Decode the chromosome and find 10 strong peaks in the recessive characteristic peak store. (2) Retention time data of 10 characteristic peak pairs were substituted into equation (1) to calculate the expansion factor \( a \) and the shift factor \( b \). (3) Input the \( P \) of the fingerprint to be tested, transform it into point set \( r \) with formula (1), and then calculate the point set: the number of peaks that can be paired with the point set: if the retention time difference between a peak in the point set \( r \) and a peak in the point set is less than the allowable threshold \( \Delta t \), the two peaks can be determined as a pair of fingerprint peaks. (4) Calculate the average retention time difference between point set \( r \) and all corresponding fingerprint peaks in point set:

\[
\Delta t = \frac{\sum \Delta t_i}{N}.
\]

Calculate individual fitness:

\[
N+1-\Delta t.
\]

2.2.3. Set Genetic Operator GA includes three basic genetic operators: selection, crossover and mutation. Firstly, the mapping from traits to genes is realized, that is, the coding work, and then the
evolutionary solution are carried out from a population representing the possible potential solution set of the problem.

2.3 Identification Fingerprint Peak
After the characteristic peak is selected, the retention time of the characteristic peak of the sample is respectively substituted into equation (1) to calculate the factor \( a \) and correct the peak of each fingerprint peak in the fingerprint to be measured.

3. Experimental Results and Discussion
This work is calculated on a 64 GB computer using MatlabR2017b software (Math Works, Inc.).

3.1. Comparison of the Performance of Different Fingerprint Peak Matching Recognition Methods
Two crude oil chromatographic chromatograms were selected as the standard fingerprint and the fingerprint to be measured, respectively, as the research example of fingerprint peak identification. From table 1, \( M_1 \) and \( M_2 \) or \( L_1 \) and \( L_2 \) groups of characteristic peaks were selected by the manual selected characteristic peak method for correction and calculation, the \( \Delta t \) obtained was significantly lower than the \( \Delta t_{\text{max}} \) obtained by the direct comparison method, indicating that the overall difference between the fingerprint to be measured and the standard fingerprint is reduced. At the same time, \( \Delta t_{\text{max}} \) was small and large, which indicated that the results obtained by the artificial selection of characteristic peaks were not reliable. The results may vary with the selection of characteristic peaks. Table 1 also showed the results obtained by the method of this paper after three calculations. The sums obtained by the three calculations were all the same, indicating that the identification results were relatively reliable. In addition, both \( \Delta t \) and \( \Delta t_{\text{max}} \) obtained by the method in this paper were smaller than the direct comparison method and the artificial selection characteristic peak method, in which \( \Delta t_{\text{max}} \) decreased more, indicating that the fingerprint peak matching recognition results obtained by the method in this paper were more accurate.

Table 1. Performance comparison of different fingerprint peak recognition methods

| Category | The unadjusted | Manually selected characteristic peak method | Manually selected characteristic peak method | GA (n=3) |
|----------|----------------|--------------------------------------------|--------------------------------------------|---------|
| \( \Delta t \) (min) | 0.2165          | 0.055                                      | 0.1003                                    | 0.0538  |
| \( \Delta t_{\text{max}} \) | 0.2774          | 0.2861                                    | 0.2095                                    | 0.1114  |

3.2. Robust Analysis of Different Fingerprint Peak Matching Identification Methods
To investigate the robustness of the method, two crude oil chromatograms containing 20 corresponding peaks were selected as the standard fingerprint and the fingerprint to be measured. Different levels of caustic noise (0.1%, 0.5%, 1.0%) were added to the retention time of the fingerprint peak of the fingerprint to be measured. In order to facilitate the calculation, \( L_1 \) and \( L_2 \) in figure 1 were manually selected as the characteristic peaks. 100 calculations were performed on each of them, and the results were presented in Table 2.
Table 2. Robustness test results of different fingerprint peak identification methods

| Noise level | GA selective characteristic peak method | Manually selected characteristic peak method |
|-------------|----------------------------------------|---------------------------------------------|
| 0.1         | 0/2000                                 | 0/2000                                     |
| 0.5         | 2/2000                                 | 20/2000                                    |
| 1.0         | 50/2000                                | 170/2000                                   |

In table 2, the top of the oblique line was the number of peaks that identify errors, and the bottom is the total number of peaks. It can be seen from table 2 that at the noise level of 0.1%, both methods are completely judged correctly. When the noise level reached 0.5%, only two fingerprint peaks were misjudged by the GA in total, while 20 fingerprint peaks were misjudged by the manually selected characteristic peak method. When the noise level reached 1%, the number of false recognition peaks of the artificial selected characteristic peak method reached 170, while the GA only has 50. The results above showed that the robustness of the proposed method was better than that of the artificial selected characteristic peak method in the case of retention time drift interference.

The relative concentration ratio distribution of four oil samples between \( n-C_9 \) and \( n-C_{40} \), as well as the basking alkanes (\( C_{19}H_{40} \)) and phytane \( C_{20}H_{42} \) was determined. Basking alkanes and phytane belonged to a group of biomarkers of petroleum, which retain the unique molecular structure of organisms when petroleum was formed millions of years ago. The relative concentration ratios of the four kinds of quantifiable oil samples were as follows:

\[
n_i = \frac{M_i}{E}
\]

Where \( n_i \) was the relative concentration ratio, \( M_i \) was the concentration of each component in the oil sample, and \( E \) was the concentration of \( n-C_{25} \).

As shown in Fig.1a and Fig.1b, the main peak carbon: Y1 was \( n-C_{32} \); Y2 for \( n-C_{31} \); Y3 for \( n-C_{16} \); Y4 for \( n-C_{14} \). Carbon distribution range: Y1 is \( n-C_{10} \sim n-C_{40} \); Y2 was \( n-C_{11} \sim n-C_{38} \); Y3 was \( n-C_{9} \sim n-C_{33} \); Y4 was \( n-C_{9} \sim n-C_{31} \). Peak shape and relative concentration ratio distribution: the relative content of the light component of Y3 (before \( n-C_{18} \)) was the highest, far higher than that of the other three oil samples. The content of light components of Y4 was also much higher than that of Y1 and Y2. The peak shapes of Y1 and Y2 were also significantly different, but both showed a distribution pattern of lower relative content of light components and higher relative content of recombination components.

![Fig.1](image1.png)

**Fig.1** Gas chromatography of (a) Oman crude oil; (b) Basra crude oil

3.3. *The Fingerprint Peak Recognition Results of the Spectrum Were Analyzed*

Fifty chromatographic fingerprints of crude oil were taken and compared in pairs. A total of 1,200 pairs of spectra were obtained, among which 27,510 actual fingerprint peaks should be identified. The absolute average value of normal alkanes in four oil samples was calculated by the criterion of normal
alkanes characteristic ratio (Table 3). This method was used to identify these fingerprint peaks by matching. The number of false peaks in 27,510 fingerprint peaks was only 8 peaks. However, with \( L_1 \) and \( L_2 \) as characteristic peaks, the manual selection method can identify 658 peaks of errors, which fully showed that the accuracy of GA algorithm was far better than that of manual selection method.

**Table 3** The absolute average value of normal alkanes in four oil samples was calculated by the criterion of normal alkanes characteristic ratio

| Characteristic ratio | Y1    | Y2    | Y3    | Y4    |
|----------------------|-------|-------|-------|-------|
| \( n-C_{17}/Pr \)    | 5.557 | 2.579 | 1.847 | 3.883 |
| \( n-C_{18}/Ph \)    | 4.727 | 2.588 | 1.804 | 3.914 |
| \( n-C_{17}/n-C_{18} \) | 0.731 | 0.764 | 1.231 | 1.713 |
| Pr/ Ph               | 0.621 | 0.767 | 1.202 | 1.727 |
| CPI                  | 0.799 | 0.926 | 0.968 | 0.962 |

**Table 4** The absolute average value of normal alkanes in four oil samples is calculated by the criterion of normal alkanes characteristic ratio

| Characteristic ratio | Y1    | Y2    | Y3    | Y4    |
|----------------------|-------|-------|-------|-------|
| \( n-C_{17}/Pr \)    | 5.557 | 2.579 | 1.847 | 3.883 |
| \( n-C_{18}/Ph \)    | 4.727 | 2.588 | 1.804 | 3.914 |
| \( n-C_{17}/n-C_{18} \) | 0.731 | 0.764 | 1.231 | 1.713 |
| Pr/ Ph               | 0.621 | 0.767 | 1.202 | 1.727 |
| CPI                  | 0.799 | 0.926 | 0.968 | 0.962 |

4. **Conclusion**

The results of simulation experiment and digital fingerprint experiment of oil spill analysis show that the proposed method of automatic matching identification of chromatographic fingerprint peaks selects characteristic peaks according to the distribution characteristics of spectral peaks of the standard fingerprint, which greatly reduces the search space of characteristic peaks and improves the computational search efficiency. Using the global optimization search capability of GA to automatically select the optimal characteristic peak from the recessive characteristic peak database can make the result of fingerprint peak matching more accurate and reliable. Therefore, it was an effective tool to calculate the similarity of chromatographic fingerprint quickly and automatically. Analysis of digital fingerprint overcomes the human error caused by the spectrogram, making the identification of oil more scientific and quantitative.

**Acknowledgments**

This work was funded by National Natural Science Foundation of China (51609131); Shandong Provincial Natural Science Foundation, China (ZR2017MEE031); Key research and development plan in Shandong Province(2019GHY112018); Shandong Provincial Higher Educational Science and Technology Foundation, China (KJ2018BBA015); Weihai Science and Technology Development Plan (2017GNS18); Shandong Jiaotong University Doctoral Research Initiation Fund; Shandong Jiaotong University “Climbing” Research Innovation Team Program(SDJTUC1802).

**Reference**

[1] R.B. Ren, K.C. Han, P.H. Zhao et al, Identification of asphalt fingerprints based on ATR-FTIR spectroscopy and principal component-linear discriminant analysis, Constr. Build. Mater. 198 (2019) 662-668.

[2] L.K. Almajmaie, O.N. Ucan, O. Bayat et al, Fingerprint recognition system based on modified multi-connect architecture (MMCA), Cogn. Syst. Res. 58 (2019) 107-113.

[3] A. Mannu, M. Ferro, G.C. Dugo et al, Improving the recycling technology of waste cooking oils: Chemical fingerprint as tool for non-biodiesel application, Waste. Manage. 96 (2019) 1-8.

[4] V. Suneel, V. Trinadha Rao, G. Suress et al, Oil pollution in the Eastern Arabian Sea from
invisible sources: A multi-technique approach, Mar. Pollut. Bull. 146 (2019) 683-695.

[5] K. Kumar, H. Singh, V. Vanit et al, Coronene diimide-based self-assembled (fibre-to-disc) fluorescent aggregates for visualization of latent fingerprints, Sensor. Actuat. B-Chem. 283 (2019) 651-658.

[6] Q. Fu, B.B. Liu, C.M. Niu, S.J. Zhao et al, The geochemical evidences of sulphur oil sources in the BZ35/36 structures of the Huanghekou Sag, Bohai Bay Basin, China, Petroleum Research. 4 (2019) 164-172.

[7] J. Orzel, B. Krakowska, I. Stanimirova et al, Detecting chemical markers to uncover counterfeit rebated excise duty diesel oil, Talanta. 204 (2019) 229-237.

[8] S. Murugesu, Q.U. Ahmed, B. F. Uzir et al, Rapid investigation of α-glucosidase inhibitory activity of Clinacanthus nutans leaf using infrared fingerprinting, Vib Spectrosc. 100 (2019) 22-29.

[9] S. Varjani, Vivek N. Upasani, Influence of abiotic factors, natural attenuation, bioaugmentation and nutrient supplementation on bioremediation of petroleum crude contaminated agricultural soil, Journal of Environmental Management. 245 (2019) 358-366.

[10] W.J. Guo, S. Zhang, G.X. Wu, Quantitative oil spill risk from offshore fields in the Bohai Sea, China, Sci. Total. Environ. 688 (2019) 494-504.

[11] I. Barra, M.A. Mansouri, Y. Cherrah et al, FTIR fingerprints associated to a PLS-DA model for rapid detection of smuggled non-compliant diesel marketed in Morocco, Vib Spectrosc. 101 (2019) 40-45.

[12] H. Besser, Y. Hamed, Causes and risk evaluation of oil and brine contamination in the Lower Cretaceous Continental Intercalaire aquifer in the Kebili region of southern Tunisia using chemical fingerprinting techniques, Environ. Pollut. 253 (2019) 412-423.

[13] V. Suneel, V. Trinadha Rao, G. Suresh et al, Oil pollution in the Eastern Arabian Sea from invisible sources: A multi-technique approach, Mar. Pollut. Bull. 146 (2019) 683-695.

[14] D.N. Vera, A.M. Jiménez-Carvelo, L. Cuadros-Rodriguez et al, Authentication of the geographical origin of extra-virgin olive oil of the Arbequina cultivar by chromatographic fingerprinting and chemometrics, Talanta. 203(2019) 194-202.

[15] V. Suneel, V. Trinadha Rao, G. Suresh et al, Oil pollution in the Eastern Arabian Sea from invisible sources: A multi-technique approach, Mar. Pollut. Bull. 146 (2019) 683-695

[16] D. Peralta, I. Triguero, S. Garcia et al, Distributed incremental fingerprint identification with reduced database penetration rate using a hierarchical classification based on feature fusion and selection, Knowl-Based Syst. 126 (2019) 91-103.