Recurrence and new cases of cancer constitute a challenging human health problem. Aquaporins (AQPs) can be expressed in many types of tumours, including the brain, breast, pancreas, colon, skin, ovaries, and lungs, and the histological grade of cancer is positively correlated with AQP expression. Therefore, the identification of aquaporins is an area to explore. Computational tools play an important role in aquaporin identification. In this research, we propose reliable, accurate and automated sequence predictor iAQPs-RF to identify AQPs. In this study, the feature extraction method was 188D (global protein sequence descriptor, GPSD). Six common classifiers, including random forest (RF), NaiveBayes (NB), support vector machine (SVM), XGBoost, logistic regression (LR) and decision tree (DT), were used for AQP classification. The classification results show that the random forest (RF) algorithm is the most suitable machine learning algorithm, and the accuracy was 97.689%. Analysis of Variance (ANOVA) was used to analyse these characteristics. Feature rank based on the ANOVA method and IFS strategy was applied to search for the optimal features. The classification results suggest that the 26th feature (neutral/hydrophobic) and 21st feature (hydrophobic) are the two most powerful and informative features that distinguish AQPs from non-AQPs. Previous studies reported that plasma membrane proteins have hydrophobic characteristics. Aquaporin subcellular localization prediction showed that all aquaporins were plasma membrane proteins with highly conserved transmembrane structures. In addition, the 3D structure of aquaporins was consistent with the localization results. Therefore, these studies confirmed that aquaporins possess hydrophobic properties. Although aquaporins are highly conserved transmembrane structures, the phylogenetic tree shows the diversity of aquaporins during evolution. The PCA showed that positive and negative samples were well separated by 54D features, indicating that the 54D feature can effectively classify aquaporins. The online prediction server is accessible at http://lab.malab.cn/~acy/iAQP.

Keywords: cancer, random forest, anova, 3D structure, machine learning
INTRODUCTION

Water, as one of the most widely existing molecules, is the basic requirement for the development of organisms. Aquaporins (AQPs) are a large and evolutionarily conserved family of proteins that facilitate water absorption and flow across cytoplasmic compartments and cell membranes in microorganisms, animals, and plants. From a previous study, aquaporins, as water channel proteins, not only take part in water molecule transport but also respond to other small molecule transport, such as glycerol, urea, ammonia, and CO2, which help those molecules cross cell membranes (Preston et al., 1992; Ma et al., 1997; Agre et al., 2002; Nielsen et al., 2002; Rojek et al., 2008). In the aquaporin family, some aquaporins are primarily water selective, such as AQP1, AQP2, AQP4, AQP5 and AQP8, while other parts of the aquaporins, such as AQP3, AQP7, AQP9, and AQP10, transport water, glycerol and other small solutes (Verkman, 2005). Aquaporins are small highly conserved membrane proteins that can selectively promote water molecule transportation through the cell membrane. Aquaporins (AQPs), with a molecular weight of 28 kDa, were first found in the membrane of human red blood cells (Agre et al., 2002). AQPs usually exist as tetramers; when water passes through these narrow channels, the conformation of AQPs can decide whether water passes through the cell membrane.

AQPs not only act as channels to take part in water and small molecule transport but are also widely related to a variety of pathophysiological statuses in cells. Evidence of AQPs in cell proliferation has aroused great interest in the research of AQPs in tumour progression (Levin and Verkman, 2006; Zhang et al., 2010; Jung et al., 2011; Nakahigashi et al., 2011; Di Giusto et al., 2012; Direito et al., 2016; De Ieso and Yool, 2018). At present, AQPs can be expressed in many types of tumours, including in the brain (Maugeri et al., 2016; Lan et al., 2017), breast (Jung et al., 2011; Lee et al., 2014; Jensen et al., 2016), skin (Hara-Chikuma and Verkman, 2008a), ovaries (Kasa et al., 2010; Jung et al., 2011; Nakahigashi et al., 2011; Di Giusto et al., 2016; Direito et al., 2016; De Ieso and Yool, 2018). At present, AQPs can be expressed in many types of tumours, including in the brain (Maugeri et al., 2016; Lan et al., 2017), breast (Jung et al., 2011), pancreas (Arsenijevic et al., 2019), colon (Nagaraju et al., 2016), skin (Hara-Chikuma and Verkman, 2008a), ovaries (Kasa et al., 2019) and lung (Chae et al., 2008). There was a positive correlation between the histological tumour grade and AQP expression, such as the expression of AQP4 in diffuse astrocytoma (Saadoun et al., 2002a; Kröger et al., 2004).

For colorectal cancer, the expression of AQP8 decreased (Fischer et al., 2001), while that of AQP1, AQP3 and AQP5 increased (Moon et al., 2003), indicating that AQPs can be expressed in tumours in humans. In general, AQP expression is upregulated in tumours. Therefore, many studies speculate that aquaporins allow water to penetrate, resulting in rapid tumour mass formation. In astrocytomas, the expression level of AQP4 is related to the amount of oedema but not to survival status (Saadoun et al., 2002a; Warth et al., 2007). Recent studies have indicated that AQPs, as prognostic markers, have a potential role in tumour-associated oedema because they participate in angiogenesis, tumour cell migration and proliferation (Saadoun et al., 2005a; Saadoun et al., 2005b; Hara-Chikuma and Verkman, 2006; Auguste et al., 2007; Hara-Chikuma and Verkman, 2008b). AQP1, AQP4 and AQP9 are expressed in brain tumours, and AQP4 expression increases with the severity of brain oedema (Saadoun et al., 2002a; Ding et al., 2010; Wang and Owler, 2011; Ding et al., 2013; Maugeri et al., 2016; Lan et al., 2017). In brain, lung, prostate and colon tumours, AQP1 with high expression participates in cell migration and tumour angiogenesis (Saadoun et al., 2002b; Saadoun et al., 2005b; Mobasher et al., 2005; Kang et al., 2008). AQP3 has increased expression in ESCA, COAD, LUAD and LIHC (Marlar et al., 2017). AQP3 knockout mice can inhibit the development of skin tumours, and tumorigenesis can utilize ATP produced by AQP3-mediated glycerol transport (Hara-Chikuma and Verkman, 2008a). AQP5 is also related to the migration, metastasis, and poor prognosis of cancer cells in BRCA (breast cancer) (Jung et al., 2011; Lee et al., 2014; Jensen et al., 2016). AQP5-regulating miRNAs inhibit BRCA cell migration through exosome-mediated delivery (Park et al., 2020). Under exosome-mediated delivery, AQP5-regulated miRNAs inhibit BRCA cell migration (Park et al., 2020).

Aquaporins play a role in the development and prognosis of various cancers, so the machine learning recognition method of aquaporins is also one of the hot spots in cancer research. Machine learning methods are applied to establish a novel and efficient classification model of aquaporins and are helpful to accelerate the recognition of aquaporins. The amino acid sequence composition of the protein is considered to be a sequence feature of the protein (Tyagi et al., 2013).

There are two methods for protein classification methods, as follows: one is based on protein sequence information (Liu et al., 2020a; Zhang et al., 2021), and the other is based on protein structure features (Liu et al., 2019; Cai et al., 2020). The sequence-based protein classification method extracts features by using the amino acid composition, amino acid number and other sequence information of the protein sequence (Liu et al., 2014). These methods are efficient and useful in predicting a large number of protein sequence datasets (Lou et al., 2014). At present, there are various studies on the classification of protein sequences, such as using logistic regression and support vector machine (SVM) methods to predict DNA binding proteins (Shen and Zou, 2020; Liu et al., 2021a) by considering amino acid proportions, amino acid compositions, amino acid spatial asymmetric distributions and biological coding characteristics of evolutionary information (Szilágyi and Skolnick, 2006; Kumar et al., 2007). The protein classification method based on protein structure identifies proteins by using structure and sequence information (Liu et al., 2014). Previous studies have focused on positive electrostatic potential, protein surface, overall charge and positive patches (Shanahan et al., 2004; Bhardwaj et al., 2005), which have achieved excellent results. Under certain conditions, the prediction accuracy of three protein motifs (helix turning helix, helix hairpin helix, and helix loop helix) is 91.1%, which indicates that this method is efficient for protein determination (Cai et al., 2009).

In our work, to promote the rapid application of AQPs in cancer treatment, a powerful sequence-based analysis method to distinguish the AQPs and cross validation was applied for results demonstration (Figure 1). It is important to develop an effective model to predict AQPs. We propose a sequence-based AQP prediction model that performs stably on various classifiers. The AQP classification model uses the 188D feature extraction.
method, applies ANOVA to reduce the dimensionality, and uses different algorithms to optimize the AQP classification model. 188D is a characteristic of the frequency of continuous amino acid residues in proteins. ANOVA is used to prune features without affecting the accuracy of the predictor.

**MATERIALS AND METHODS**

**Dataset**

A high-quality dataset is essential for reliable and accurate predictor building (Su et al., 2021). Aquaporin was taken as the positive sample, and the protein sequence was collected from the protein database of the UniProt website (https://www.uniprot.org/) (Chen et al., 2016). Negative samples such as nonaquaporins were extracted from the Pfam database (http://pfam.xfam.org/). To ensure the reliability of the aquaporin dataset, we applied the following criteria to optimize the data: first, the sequences annotated as “prediction” were eliminated; second, we deleted the sequences of other protein fragments; through screening steps, 239 aquaporin sequences and 10,713 nonaquaporin sequences were obtained; third, the CD-HIT program (Fu et al., 2012) was used to eliminate redundant sequences and to avoid overestimating the prediction model (Zou et al., 2020). The cut-off of sequence identity is set to 90%. Finally, 151 aquaporins and 8,994 nonaquaporins were obtained to form the final dataset.

**Features Extraction**

One of the main factors for the performance accuracy of the prediction model is the quality of sample feature extraction. The prediction of the protein model mainly depends on the coding strategy of the protein sequence. According to the coding strategy of the protein sequence, the amino acid sequence can be transformed into a numerical vector (Liu et al., 2019; Muhammod et al., 2019; Zhu et al., 2019; Chen et al., 2020; Fu et al., 2020; Tang et al., 2020; Wang et al., 2020; Shao et al., 2021). In this paper, the global protein sequence descriptor (GPSD) method was used to represent the amino acid sequence. Global protein sequence descriptor (GPSD), known as 188 days method. This method mainly converts the sequence into a numerical vector according to the amino acid properties in the protein sequence and generates 188 features. These 188D features contain the information and properties of amino acid sequences [48,49]. According to the description of the GPSD method, the 188D features can be divided into two parts. The first part is the composition of amino acids. The first 20D features were obtained by calculating the frequency of amino acids in the protein sequence. The second part is to calculate the physicochemical properties of amino acids, which constitute 168 characteristics. Previous studies have provided detailed information on the eight physicochemical properties of amino acids (Lin et al., 2013; Liu et al., 2018; Li et al., 2019a). The protein sequence was encoded by CTD (C: composition, t: transition, D: distribution) mode to generate 21D features. Three groups were generated for 20 amino acids for each property. C is the occurrence frequencies (1 × 3D = 3D). T is the transition frequency (1 × 3D = 3D). D is the first, 25, 50, 75% and last position of a certain group in the peptide sequence (5 × 3 = 15D). Therefore, 8 * (3 + 3 + 15) = 168 features were produced for the CTD model.

**Classifier**

To find the most suitable machine learning algorithm, six commonly used classifiers are applied, including random forest (RF) (Ru et al., 2019), NaiveBayes, support vector machine (SVM) (Wei et al., 2018a; Dao et al., 2020a; Dao et al., 2020b; Wei et al., 2020), XGBoost (Yu et al., 2021a; Yang et al., 2021), logistic regression (LR) and decision tree (DT) (Li et al., 2019b). These efficient machine learning algorithms are usually used for feature analysis.

**Feature Selection**

For machine learning model building, features extracted from sequences always contain noise. A feature selection strategy to solve the information redundancy and overfitting problem can improve the feature representation ability (He et al., 2021). Analysis of variance (ANOVA) (Blanca et al., 2017; Wei et al., 2018b; Tang et al., 2018; Su et al., 2019a; Jung et al., 2019; Su et al., 2020; Liu et al., 2021b; Jin et al., 2021) has been used to analyse these characteristics and has been widely used in RNA, DNA and protein prediction. In this study, ANOVA is used to select the optimal features for model training. The feature subset with low redundancy is selected by ANOVA. We sort the original features.
based on the ANOVA feature sorting algorithm and apply the IFS strategy to search the optimal feature subset.

Performance Standard
To evaluate the prediction accuracy of the model, the data of the following four formulas are usually used to solve the problem of classification prediction.

\[
Acc = \frac{TP + TN}{(TP + TN + FP + FN)}
\]

\[
Sn = \frac{TP}{(TP + FN)} \quad Sp = \frac{TN}{(TN + FP)}
\]

\[
MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]

Accuracy (Acc), specificity (Sp), sensitivity (Sn) and Matthew correlation coefficient (MCC) were the commonly used evaluation parameters (Jiang et al., 2013; Wei et al., 2014; Wei et al., 2017a; Wei et al., 2017b; Manavalan et al., 2019a; Manavalan et al., 2019b; Su et al., 2019b; Hong et al., 2019; Zeng et al., 2019; Zhang et al., 2020a; Ding et al., 2020b; Hasan et al., 2020; Huang et al., 2020; Jin et al., 2021; Shao and Liu, 2021; Zhu et al., 2021). In the formulas, TP was the true positive number, TN was the true negative number, FP was the false-positive number and FN was the false negative number.

A receiver operating characteristic (ROC) curve was applied to study the prediction performance of the model. The area under the ROC curve (AUC) was used to assess the prediction performance of the model. AUC values of 0.5 and one represent random and perfect models, respectively (Zeng et al., 2017; Zeng et al., 2018; Dao et al., 2019; Fong et al., 2019; Lai et al., 2019; Lin et al., 2019; Zhu et al., 2019; Zhang et al., 2020b; Charoenkwan et al., 2020; Ding et al., 2020a; Ding et al., 2020b; Hasan et al., 2020; Huang et al., 2020; Jin et al., 2020; Li et al., 2020; Wang et al., 2020; Wu and Yu, 2021).

Construction of 3D Structure for AQP
To verify the localization of aquaporins, the website (http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc-2/) of the protein localization website and transmembrane prediction website (https://www.novopro.cn/tools/tmhmhm.html) were applied to predict the subcellular localization and transmembrane structure of aquaporins. At the same time, Phyre2 software (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) was applied for the 3D structure prediction of aquaporins. The prediction results were visualized by PyMOL (version 2.5.1) software (https://pymol.org/2/).

Construction of Aquaporins Phylogenetic Tree
The phylogenetic tree of aquaporins was constructed to analyse the evolutionary diversity of the protein. Aquaporin sequence alignment results were analysed by MAFFT online software (https://mafft.cbrc.jp/alignment/server/) and used to construct a phylogenetic tree using IQ-TREE software (multicore version 1.6.12). The best fitting model for the phylogenetic tree was LG + F + R6 (Kalyaanamoorthy et al., 2017). The ultrafast bootstrap method was used for phylogenetic assessment, and 1,000 replicates per method were chosen in this work (Guindon et al., 2010; Minh et al., 2013; Hoang et al., 2018). The tree file was visualized by the iTOL website (https://itol.embl.de/).

| TABLE 1 | Preliminary results of different feature descriptors using different classifiers. |
| 188D_P: N = 1:1 | Sn | Sp | Acc | MCC | AUROC |
| XGBoost | 98 | 96.04 | 97.033 | 0.9416 | 0.9949 |
| NaiveBayes | 98.666 | 96.041 | 96.366 | 0.9279 | 0.9763 |
| LR | 97.999 | 94.748 | 96.377 | 0.9289 | 0.9867 |
| DecisionTree | 95.999 | 93.747 | 95.689 | 0.9155 | 0.9669 |
| RF | 98.666 | 97.707 | 97.689 | 0.9544 | 0.9987 |
| SVM | 95.332 | 96.041 | 95.7 | 0.9153 | 0.9917 |

| 188D_P: N = 1:2 | Sn | Sp | Acc | MCC | AUROC |
| NaiveBayes | 98 | 97.667 | 97.793 | 0.9531 | 0.9765 |
| SVM | 90.75 | 98.334 | 96.471 | 0.9224 | 0.9965 |
| LR | 96.083 | 98.001 | 97.359 | 0.9425 | 0.9958 |
| RF | 97.332 | 98.344 | 98.012 | 0.9564 | 0.9978 |
| XGBoost | 96.708 | 96.677 | 96.682 | 0.9284 | 0.9964 |
| DecisionTree | 92.041 | 93.687 | 93.141 | 0.8518 | 0.9296 |

| 188D_P: N = 1:3 | Sn | Sp | Acc | MCC | AUROC |
| NaiveBayes | 97.333 | 96.015 | 96.357 | 0.9102 | 0.9765 |
| SVM | 90.75 | 98.334 | 96.471 | 0.9224 | 0.9965 |
| LR | 95.417 | 98.445 | 97.882 | 0.9394 | 0.9952 |
| RF | 96.666 | 98.455 | 98.013 | 0.948 | 0.9979 |
| XGBoost | 95.374 | 98.011 | 97.343 | 0.9306 | 0.995 |
| DecisionTree | 93.999 | 96.697 | 96.024 | 0.8974 | 0.9835 |

| 188D_P: N = 1:4 | Sn | Sp | Acc | MCC | AUROC |
| NaiveBayes | 97.333 | 97.18 | 97.22 | 0.9206 | 0.9771 |
| SVM | 92.083 | 99.005 | 97.617 | 0.9256 | 0.9946 |
| LR | 93.457 | 97.844 | 96.953 | 0.9074 | 0.9942 |
| RF | 94.709 | 99.166 | 98.274 | 0.9461 | 0.9967 |
| XGBoost | 95.333 | 98.668 | 98.007 | 0.9391 | 0.9958 |
| DecisionTree | 92.708 | 98.841 | 97.614 | 0.9248 | 0.9978 |

| 188D_P: N = 1:5 | Sn | Sp | Acc | MCC | AUROC |
| NaiveBayes | 96.666 | 97.084 | 97.019 | 0.9022 | 0.9773 |
| SVM | 92.083 | 99.205 | 98.015 | 0.9292 | 0.9954 |
| LR | 93.999 | 98.143 | 97.46 | 0.9136 | 0.996 |
| RF | 94.667 | 99.338 | 98.562 | 0.9486 | 0.9975 |
| XGBoost | 95.333 | 98.94 | 98.341 | 0.9414 | 0.9963 |
| DecisionTree | 91.374 | 98.01 | 96.905 | 0.8924 | 0.9469 |

| 188D_P: N (151:8,994) | Sn | Sp | Acc | MCC | AUROC |
| XGBoost | 86.084 | 99.934 | 99.703 | 0.9062 | 0.9989 |
| NaiveBayes | 96.666 | 97.977 | 97.955 | 0.9622 | 0.9793 |
| LR | 94.082 | 99.635 | 99.374 | 0.8158 | 0.9975 |
| DecisionTree | 72.208 | 99.365 | 98.918 | 0.6827 | 0.8579 |
| RF | 82.75 | 99.912 | 99.626 | 0.879 | 0.995 |
| SVM | 31.167 | 100 | 98.866 | 0.5503 | 0.9916 |

EXPERIMENT
Performance of Features Based on the 188-Dimensional Method (GPSD)
To select the best classifier for the AQP sequences, six widely used machine learning classifiers were employed to classify the features...
The results of all classification based on machine learning algorithms have similar accuracies lower than the random forest (RF) algorithm. The NaiveBayes, LR, DecisionTree and SVM algorithms have accuracies of 0.9949, 0.9763, 0.9857, 0.9569 and 0.9917, respectively. In this study, six widely used classifiers are used for classification. The ROC of the RF classifier is 0.9985, which is relatively high. In general, regarding the evaluated accuracy of the AUC, Acc and MCC values, RF had the best performance in the AQP sequence classification results and was selected as the best classifier for model building.

Effect of Feature Selection Technologies

For the AQP sequences (P: N = 1:1), random forest (RF) was the best algorithm, with the highest accuracy for the features extracted by the 188-dimensional method (GPSD). For feature extraction by the 188-dimensional method (GPSD), we applied different ratios for the number of positive and negative samples (1:1, 1:2, 1:3, 1:4, 1:5, and 151:8994), and the results were classified by six machine learning classifiers (XGBoost, NaiveBayes, LR, decision tree, RF and SVM). The results of all classifiers in the tenfold cross-validation were compared, and the comparison results are shown in Table 1.

The results of Table 1 show that the different proportions of positive and negative samples indicated that P: N = 1:1 was the best ratio for the following analysis. Although the values of 1:2, 1:3, 1:4, 1:5 and 151:8998 have higher values in SP and ACC, the values of Sn, MCC and AUROC are lower compared with P: N = 1:1. The increase in negative samples causes data imbalance and overfitting of the model. Therefore, the positive and negative sample ratio column of P: N = 1:1 is selected for model building.

For the AQP sequences (P: N = 1:1), random forest (RF) was the best algorithm, with the highest accuracy for the features extracted by the 188-dimensional method (GPSD) (AUC = 0.9985), ACC = 97.689, MCC = 0.9544, Sn = 98.666, Sp = 96.04. XGBoost is the second algorithm with a slightly lower accuracy (AUC = 0.9949, Acc = 97.033, MCC = 0.9416, Sn = 98, Sp = 96.04) compared with the random forest (RF) algorithm. The NaiveBayes, LR, DecisionTree and SVM algorithms have similar accuracies lower than the random forest (RF) algorithm for AQP sequence classification based on the 188-dimensional method (GPSD). The results in Figure 2 indicated that RF was the best classifier with an accuracy of 0.9985, while the other classifiers of XGBoost, Naivebayes, LR, decision tree and SVM had accuracies of 0.9949, 0.9763, 0.9857, 0.9569 and 0.9917, respectively. In this study, six widely used feature selection methods based on ANOVA are shown in Table 2. In addition, the optimal feature 54D is selected by combining ANOVA with an incremental feature selection (IFS) strategy, as shown in Figure 3A. The comparison results show that the accuracy of the optimal feature selected (ACC = 97.689) is slightly higher than that of the original feature (ACC = 97.356) (Table 2). Therefore, the ANOVA feature selection method was selected for feature optimization.

The PCA method was used to visually analyse the optimal feature (54D) after feature selection by the feature selection method (Figure 3B). Figure 3B indicates that positive and negative samples can almost be separated in the two-dimensional visualization diagram, which indicates that the 54D feature can effectively classify AQP proteins.

Feature Distribution Analysis

In this study, we performed feature analysis after feature selection. By analysing these 188D features, we determine the attribute information contained in these features. The results of feature analysis are shown in Figure 3C. According to the best feature analysis of the F-score value obtained by ANOVA, the features with an F-score value greater than 100 have a greater contribution to the classification. It can be seen from the figure that among the 188D features, the first is the 26th dimension feature, which is neutral/hydrophobic, followed by the 21st dimension feature, which is hydrophobic. The 26th dimension feature (neutral/hydrophobic) and 21st dimension feature (hydrophobic) signs showed that AQPs contained hydrophobic amino acids, which may be associated with the structural and functional properties of AQPs.

Structure Analysis of AQPs

Through feature selection, we know that hydrophobic features (the 26th dimension feature and 21st dimension feature) are the...
FIGURE 3 | Two-step feature selection result display (A) 10-fold CV and independent test accuracy of the RF classifier with the feature number varied (B) dimension reduction results based on the PCA method for the original data with a total of 188 dimensions (C) feature ranking of the F-score method obtained by ANOVA for the data with 188 features.

FIGURE 4 | The structure of AQPs (A) The prediction distribution of the transmembrane structure for AQP6_HUMAN (B) model of the structure of an AQP showing the principal features of the protein, NPA: asparagine-proline-alanine motifs; M1–M6: the transmembrane structure (C–E) 3D constructure of AQP6_HUMAN, AtPIP1-4 and AQPZ-ECLOI.
most significant features and make a great contribution to classification. Therefore, we analysed the protein localization of the AQP protein sequence, and the results showed that all AQP proteins were located on the cell membrane (Supplement Table 1). Cells are distinguished by a thin membrane. The core of the membrane is hydrophobic, which means it repels water. Many signals and nutrients cannot pass through the membrane itself but can pass through proteins across the membrane. Membrane proteins are essential for living cells, and plasma membrane proteins also have properties such as hydrophobicity, low solubility and low abundance. Therefore, the enrichment and classification extraction methods of soluble proteins cannot be used for plasma membrane proteins, mainly because the expression level of plasma membrane proteins in cells is very low, and they are highly hydrophobic in nature, which makes them easier to precipitate in aqueous solution and difficult to extract (Luche et al., 2003; Rawlings, 2016).

The Phyre2 website was used to analyse the transmembrane structure of HmAQP7. Figure 2 shows that there are six α-helix transmembrane domains (Figure 4A): M1, M2, m3, M4, M5 and M6 (Figure 4B). A six-α-helix transmembrane domain forms a pore on the cell membrane to supply water molecules through the cell membrane. When the AQP protein folds, loops B (HB) and E (HE), which retain the lipophilic half helix, project to the protein molecular centre, making the highly conserved Asn-Pro-Asp (NPA) motif present the opposite direction, thus regulating the single file conductance of water and acting as a cation and proton exclusion filter (Figures 4B–E).

### Evolution and Diversity

Aquaporin is a conserved membrane protein that contains highly conserved NAP domains and α-helical transmembrane domains in bacteria (Figure 4E), plants (Figure 4D) and humans (Figure 4C). To better verify the phylogenetic and evolutionary relationship of AQPs, 151 AQP protein sequences containing human, mouse, insect, fungus and bacteria were applied to construct a phylogenetic tree (Figure 5).

The results indicated that the 151 AQP protein sequences were divided into eight groups (Figure 5). The length of branches indicates the genetic relationship of AQP sequences. Among them, group III and group IV belong to plant and bacteria branches, respectively. Group II is the most complex branch, including the aquaporins of fungi, bacteria and animals. Among them, the VIa and VIIIa branches are plant subfamilies. AQPs of the VIIa and VIIb subfamilies belong to animals and insects, respectively. Group V contains one bacterial AQPZ and 15 animal AQPs, of which 7 belong to Tardigrade.

### Expression of AQPs in Tumour Tissue

AQPs are considered to be important prognostic markers of cancers (Chow et al., 2020), so the expression of AQPs in cancer tissues is also crucial. Figure 6 shows the expression level of AQP transcripts in 33 tumour tissues. AQP1_HUMAN has a high expression level in all tumour tissues and plays an important role in tumour angiogenesis and endothelial cell migration (Saadoun et al., 2005b). AQP3_HUMAN is expressed in almost all tumour tissues except ACC, LGG, UVM and AQP3_HUMAN-mediated glycerol transport, which allows the production of ATP for tumorigenesis. AQP3_HUMAN knockout mice can be resistant to carcinogen induction skin tumours (Hara-Chikuma and Verkman, 2008a). AQP3_HUMAN and AQP5_HUMAN were also expressed in COAD (Moon et al., 2003), while AQP5_HUMAN expression in human COAD is related to cell proliferation and metastasis. In BRCA, AQP5_HUMAN overexpression is associated with (Jung et al., 2011; Lee et al., 2014; Jensen et al., 2016) migration and poor prognosis in BRCA patients. Consistently, AQP5_HUMAN regulates miRNA migration through exosome-mediated (Park et al., 2020) and inhibits BRCA cell migration. AQP2_HUMAN, AQP12A_HUMAN, AQP12B_HUMAN and MIP had low expression levels in 33 tumour tissues, AQP4_HUMAN was highly expressed in GBM and LGG, and AQP9_HUMAN was highly expressed in LIHC.

### Web Server Implementation

To facilitate the prediction of aquaporins, a user-friendly online server named iAQPs-RF is applied, which can be accessed from http://lab.malab.cn/~acy/iAQP. The protein sequences (FASTA format) were identified to determine whether aquaporins or non-aquaporins use the web server by users. First, the FASTA format protein sequences are entered or pasted in the left blank box and the submit button is clicked; finally, the results are displayed on the right box. If you want to restart a new task, a clear button or the resubmit button was clicked to clear the sequences in the input box. Finally, new query protein sequences were allowed to enter the input box. The home page provides links of the contact information of authors and relevant data to download.
CONCLUSION

The accurate identification of aquaporins by iAQPs can greatly promote the prediction of aquaporins and research on tumour diseases. In this study, we used the GPSD method to extract protein sequence features and the optimal random forest algorithm to construct new computational aquaporin identifier iAQPs-RF. Combined with the feature selection technique ANOVA, 54 optimal features are selected to build the predictor. According to the F-score value obtained by ANOVA, the 26th dimension feature and 21st dimension feature are ranked as the first and second dimension features among the 188 days features, respectively, and these two features possess neutral/hydrophobic characteristics. These two dimensional features make a great contribution to the classification of aquaporins. At the same time, through the location and 3D structure prediction of aquaporins protein, although the protein divided into eight groups and has diversity in evolution, all the proteins belong to plasma membrane proteins, and the protein sequence contains six α-helix transmembrane domains. The membrane proteins are hydrophobic and contain many hydrophobic amino acids (Luche et al., 2003; Rawlings, 2016), so these results are consistent with aquaporin classification.

The best CV evaluation accuracy of iAQPs-RF was 97.689%. At the same time, a network server is established. iAQPs-RF are expected to be a robust and reliable tool for aquaporin identification. Future work will focus on exploring deep learning to improve the performance of the model.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

LX, XS and LZ designed the research; ZC and SJ performed the research; ZC and DZ analyzed the data; ZC wrote the manuscript. All authors read and approved the manuscript. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

FUNDING

The work was supported by the National Natural Science Foundation of China (No.62002244, No.62001311) and the Post-doctoral Foundation Project of Shenzhen Polytechnic China (No. 6020330002K).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2022.845622/full#supplementary-material

REFERENCES

Agre, P., King, L. S., Yasui, M., Guggino, W. B., Ottersen, O. P., Fujiyoshi, Y., et al. (2002). Aquaporin Water Channels - from Atomic Structure to Clinical Medicine. J. Physiol. 542, 3–16. doi:10.1113/jphysiol.2002. 020818

Arsenijevic, T., Perret, J., Van Laethem, J.-L., and Delporte, C. (2019). Aquaporins Involvement in Pancreas Physiology and in Pancreatic Diseases. Ijms 20 (20), 5052. doi:10.3390/ijms20205052

Auguste, K. I., Jin, S., Uchida, K., Yan, D., Manley, G. T., Papadopoulos, M. C., et al. (2007). Greatly Impaired Migration of Implanted Aquaporin-4-deficient Astroglial Cells in Mouse Brain toward a Site of Injury. FASEB j. 21 (1), 108–116. doi:10.1096/fj.06-6848com
